BEYOND CTT AND IRT: USING AN INTERACTIONAL MEASUREMENT MODEL TO INVESTIGATE THE DECISION MAKING PROCESS OF EPT ESSAY RATERS

BY

DIANA XIN WANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Professor Frederick Davidson, Chair
Associate Professor Kiel Christianson
Professor Hua-Hua Chang
Associate Professor Randall Sadler

ABSTRACT

The current study as a doctorate dissertation investigates the gap between the nature of ESL performance tests and score-based analysis tools used in the field of language testing. The purpose of this study is hence to propose a new testing model and a new experiment instrument to examine test validity and reliability through rater's decision making process in an ESL writing performance test.

A writing test as a language performance assessment is a multifaceted entity that involves the interaction of various stakeholders, among whom essay raters have a great impact on essay scores due to their subjective scoring decision, hence influencing the test validity and reliability (Huot, 1990; Lumley, 2002). This understanding puts forward the demand on the development and facilitation of methodological tools to quantify rater decision making process and the interaction between rater and other stakeholders in a language test. Previous studies within the framework of Classic Testing Theory (CTT) and Item Response Theory (IRT) mainly focus on the final outcome of rating or the retrospective survey data and/or rater's think-aloud protocols. Due to the limitation of experimental tools, very few studies, if any, have directly examined the moment-to-moment process about how essay raters reach their scoring decisions and the interaction per se.

The present study proposes a behavioral model for writing performance tests, which investigates raters' scoring behavior and their reading comprehension as combined with the final essay score. Though the focus of this study is writing assessment, the current research methodology is applicable to the field of performance-based testing in general. The present framework considers the process of a language test as the interaction between test developer, test taker, test rater and other test stakeholders. In the current study focusing on writing performance

test, the interaction between test developer and test taker is realized directly through test prompt and indirectly through test score; on the other hand, the interaction between test taker and test rater is reflected in the writing response. This model defines and explores rater reliability and test validity via the interaction between text (essays written by test-takers) and essay rater. Instead of indirectly approaching the success of such an interaction through the final score, this new testing model directly measures and examines the success of rater behaviors with regard to their essay reading and score decision making. Bearing the "interactional" nature of a performance test, this new model is named as the Interactional Testing Model (ITM).

In order to examine the online evidence of rater decision making, a computer-based interface was designed for this study to automatically collect the time-by-location information of raters' reading patterns, their text comprehension and other scoring events. Three groups of variables representing essay features and raters' dynamic scoring process were measured by the rating interface: 1) Reading pattern. Related variables include raters' reading rate, raters' go-back rate within and across paragraphs, and the time-by-location information of raters' sentence selection. 2) Raters' reading comprehension and scoring behaviors. Variables include the time-by-location information of raters' verbatim annotation, the time-by-location information of raters' comments, essay score assignment, and their answers to survey questions. 3) Essay features. The experiment essays will be processed and analyzed by Python and SAS with regard to following variables: a) word frequency, b) essay length, c) total number of subject-verb mismatch as the indicator of syntactic anomaly, d) total number of clauses and sentence length as the indicators of syntactic complexity, e) total number and location of inconsistent anaphoric referent as the indicator of discourse incoherence, and f) density and word frequency of sentence

connectors as indicators of discourse coherence. The relation between these variables and raters' decision making were investigated both qualitatively and quantitatively.

Results from the current study are categorized to address the following themes:

1) Rater reliability: The rater difference occurred not only in their score assignment, but also in raters' text reading and scoring focus. Results of inter-rater reliability coincided with findings from raters' reading time and their reading pattern. Those raters who had a high reading rate and low reading digression rate were less reliable.

2) Test validity: Rater attention was assigned unevenly across an essay and concentrated on essay features associated to "Idea Development". Raters' sentence annotation and scoring comments also demonstrated a common focus on this scoring dimension.

3) Rater decision making: Most raters demonstrated a linear reading pattern during their text reading and essay grading. A rater-text interaction has been observed in the current study. Raters' reading time and essay score were strongly correlated with certain essay features. A difference between trained rater and untrained rater was observed. Untrained raters tend to over emphasis the importance of "grammar and lexical choice".

As a descriptive framework in the study of rating, the new measurement model bears both practical and theoretical significance. On the practical side, this model may shed light on the development of the following research domains: 1) Rating validity and rater reliability. In addition to looking at raters' final score assignments, IRM provides a quality control tool to ensure that a rater follows rating rubrics and assigns test scores in a consistent manner; 2) Electronic essay grading. Results from this study may provide helpful information to the design and validation of an automated rating engine in writing assessment. On the theoretical side, as a supplementary model to IRT and CTT, this model may enable researchers to go beyond simple

post hoc analysis of test score and get a deeper understanding of raters' decision making process in the context of a writing test.

ACKNOWLEDGEMENTS

I would like to acknowledge and thank each member of my committee. I would never have been able to finish my dissertation without their encouragement, support, and advice. I would like to express my deepest gratitude to my advisor, Dr. Fred Davidson, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I cannot thank him enough for his invaluable guidance and constant support at every stage. I would also like to thank the rest of my committee members, Dr. Kiel Christianson, Dr. Hua-Hua Chang, and Dr. Randall Sadler, for guiding my research for the past several years and helping me to develop my background in psychology, measurement, and educational technology. My research would not have been possible without their helps.

While there are many fellow doctoral students who become friends during the past few years, two of them, Chih-Kai Lin and Sun Joo Chung, need a special thank you for their invaluable feedback on my dissertation. I would also like to thank the graduate students for participating in my study and providing valuable inputs on the EPT writing test.

Finally, I wholehearted thank my family for always supporting me and encouraging me with their best wishes. I am forever grateful to my parents who always have faith in me and trust me in everything I do and in every dream that I want to pursue. Yudong, a special thank you goes to you for your unconditional support in my personal life and professional growth.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 The Application of Measurement Theories in Language Testing

Since 1980s, the field of language testing has emerged and developed under the influence of two major disciplines --- applied linguistics and measurement theories. The study of language testing first took place within the framework of applied linguistics. The development in language testing has incorporated advances in the research areas of language acquisition, language teaching, and the study of language proficiency. From this perspective, the purpose of language testing study is to construct a theoretical framework of language ability and provide a means to describe and assess, based on a given norm of target language, the language ability of individuals at a certain stage of development (Bachman, 1998). With the application of measurement theories to the assessment of language ability, language test can be defined as a measurement instrument explicitly designed to elicit a specific set of language sample from test takers' behavior response, which then will be graded according to prescribed measurement norms.

1.1.1 Theoretical Advances in Language Testing

In language testing, the essential task is to determine the nature of language ability, or language proficiency. The concept of language proficiency was originally derived from views of structuralist linguistics which consider language as a composition of discrete components and skills, and also from a psychological view in which ability is treated as a unidimensional attribute (Bachman, 1998). Based on recent findings, language testers have reached the general consensus that language proficiency consists of a number of distinct but interrelated component abilities.

As one of these language competences, the concept of communicative competence was originally proposed by Hymes (1972). This notion was further developed in the early 1980s by Canale and Swain, who defined the communicative competence as, rather than the knowledge of language itself, an individual's underlying systems of knowledge and skill required for communication (Canale & Swain, 1980). Canale (1983) proposed that there are four components of communicative competence which include grammatical competence, sociocultural competence, discourse competence and strategic competence.

Based on this componential view of communicative competence, Bachman (1991) put forward a "multi-componential" view of language proficiency, in which he expanded Swain and Canale's description of communicative language ability by acknowledging the role of strategic competence (Bachman, 1990, 1991; 2000). Bachman's model is also called "interactional model" in which language proficiency, rather than a unitary trait, is multicomponential which consists of "a number of interrelated specific abilities as well as a general ability or set of general strategies or procedures" (1990, p. 673). In this model, test takers' language ability includes language knowledge and metacognitive strategies, and the test method includes characteristics of the environment, rubric, input, expected response and the relationship between input and expected response (Bachman, 1990; Bachman & Palmer, 1996). McNamara (1995) further expanded Bachman's model by adding the social dimension of language proficiency. He also pointed out that rater's perception of test taker's performance and rater's use of rating scales are potential influences on test score.

The theoretical change in language testing also affects the advances in test development and research methodology. The multicomponential model, in particular, puts forward a demand on appropriate tools to examine the interactive aspects in language testing.

1.1.2 Measurement Theories in Language Testing

The methodological development in measurement theory provides powerful tools for the study of language testing. Two measurement theories including Classical Test Theory (CTT) and Item Response Theory (IRT) have been widely accepted as the major measurement frameworks for the construction and interpretation of language tests with two complementary objectives—reliability and validity (Bachman, 1990).

1.1.2.1 Classical Test Theory

The CTT framework is based on the assumption that the observed score for an individual can be viewed as the sum of two components: a true score and a random error (Lord & Novick, 1968). In CTT, the degree to which the true score accounts for the variance in observed score is defined as test reliability, which represents the accuracy with which a test linearly ranks a group of test-takers (Lord & Novick, 1968; Mislevy, 1993). In the 1970s and early 1980s, CTT was the primary psychometric tool to estimate language abilities in reliability research. Before the emergence of IRT in language testing in the 1970s, CTT had also dominated the area of standardized testing. The major theoretical advantage of CTT is that it is built on relatively weak theoretical assumptions, which makes it easy to apply in various testing situations (Hambleton & Jones, 1993).

Nevertheless, CTT has its own limitations when applied in language tests. A primary criticism is related to the instability of item statistics and person statistics produced by CTT as well as the circular dependency between them. It is believed that the item statistics are person sample dependent and person statistics are item sample dependent. This circular dependency poses some theoretical difficulties in the application of CTT in measurement situations including

test equating and adaptive testing (Hambleton & Swaminathan, 1985). Another problem of CTT is that the definition of error and subsequent reliability coefficients vary across different reliability estimates, hence reliability indices in CTT consider only one source of measurement error at a time. Therefore, it is difficult to make a decision when several reliability coefficients differ substantially (Hambleton, 1989).

In order to overcome these limitations of the old version CTT model, Generalizability theory (G-theory), as an extension of CTT, was originally developed by Crobach and his colleagues to account for the dependability of a behavioral measurement (Cronbach, Rajaratnam, & Gleser, 1963; Gleser, Cronbach, & Rajaratnam, 1965; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). G-theory heavily roots in the basic idea of variance decomposition of a person-by-item response matrix (Hoyt, 1941), the framework of factorial design, and also the theory of "domain sampling" explicitly developed by Tryon (1957). In G-theory, an observed score is considered as a sample from a hypothetical universe of generalization, which is a domain of uses and/or abilities to which the test scores are to be generalized. Therefore, the interpretation of a test score represents the generalization from a single measure to a universe of measure. Reliability in G-theory is a matter of how accurately the observed score allows generalization concerning a person's ability to a universe of defined situations.

G-theory was introduced into language testing in 1982 (Bolus, Hinofotis, & Bailey, 1982). Since then, this theory has gradually gained its popularity in various domains of language testing such as the consistencies across items, subtests and languages (Brown, 1999; also see Bachman, 2000), the agreement in placement decisions (Kunnan, 1992) and factors that affect rater reliability (Shohamy, Gordon, & Kraemer, 1992; Lynch & McNamara, 1998). A big advantage of G-theory is that it enables test specialists to investigate simultaneously multiple

sources of measurement error and the interactions. G-theory is more powerful than the traditional CTT model in estimating the effect of the number of items and raters, thus helping test developers to maximize the test reliability within a given administration context (Bachman, 1990). In addition, this model provides comparable reliability estimates for language tests that differ in test length and number of rater (Bachman, 1990).

Despite its advantages, G-theory has been criticized from several perspectives. Rozeboom (1978) questioned the conceptual existence of a domain or a universe of generalization. He pointed out that it is logically impossible to sample from a domain in order to make the assumptions necessary to generate both coefficient alpha and G-theory. Other limitations of G-theory include the lack of a basis for determining how a person might respond to a particular item, the difficulty in comparing the performance of persons who take different forms of an assessment, and the lack of procedures for determining how measurement error varies across the levels of the construct under investigation (Smith & Kulikowich, 2004). Further more, the CTT models including G-theory fail to predict how an individual test taker responds to a given test item (see Bachman, 1990).

1.1.2.2 Item Response Theory

Theoretically, IRT overcomes the major weakness of CTT, which is the circular dependency of item/person statistics. As a result, in theory, IRT models produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered. This “invariance” property of item and person statistics of IRT has been illustrated theoretically (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991) and has been accepted within the measurement community. Since the beginning of the 1970s,

IRT has gradually replaced the dominating role of CTT and has become a very important measurement framework (Hambleton, Swaminathan & Rogers, 1991). Being more theory grounded, IRT models the probabilistic distribution of examinees' success at the item level, which is in contrast to CTT's primary focus on test-level information. One major assumption of IRT is that the response to any item is unrelated to any other items at the same trait level. In addition, the latent ability of a test taker is independent of the content of a test. The relationship between the probability of answering an item correctly and the ability of a test taker can be modeled in different ways depending on the nature of the test (Hambleton, Swaminathan & Rogers, 1991). Another important assumption is the appropriate dimensionality, which means that IRT contains the right number of trait level estimates per person for the data. In the current IRT models, unidimensionality is a common assumption which indicates that items in a test measure a single latent ability. IRT models also follow the assumption that it does not matter which items are used in order to estimate the test-takers' ability. This assumption makes it possible to compare test takers' result despite the fact that they took different versions of a test (Hambleton & Swaminathan, 1985).

The conceptualization of IRT opens the door to solving many practical problems in language testing. First, it allows the estimate of item statistics and the abilities of test takers so that they are not sample dependent for large-scale standardized language proficiency tests (Bachman & Eignor, 1997; Pollitt, 1997). The application of IRT has also brought great advances in computer-adaptive language testing which selects the best item for an examinee based on the information provided by available items and the examinee's proficiency estimate, thus making language tests more efficient and adaptable to individual test takers (e.g. Tung, 1986).

In research of language testing, several different IRT models are incorporated already, among which the Rasch model (one-parameter IRT model) remain the most widely used (e.g. de Jong, 1986; Adams, Griffin, & Martin, 1987; Lynch, Davidson, & Henning, 1988; McNamara, 1991; Bolt, 1992; as cited in Bachman, 1990). A multifaceted version of the Rasch measurement (FACETS) model for ordered response categories was developed by Linacre (1989). FACETS has been applied to investigate the effects of raters and tasks, or other multiple measurement facets in language performance assessments (Bachman, Lynch & Mason, 1995; Brown, 1995; Lumley & McNamara, 1995; Weigle, 1998).

Despite the advantages of IRT over the CTT-based methods, this model has its own limitations. Henning (1991) argued that problems might be encountered in the use of IRT with the validity of item banking techniques in language testing settings. Another major limitation of IRT is that a large number of examinees must be tested before it reaches stable and reliable application. In testing practices, generally speaking, IRT is thus more applicable for full item analysis when the numbers of students being tested are very large (Hulin, Drasgow, & Parsons, 1983).

In summary, classical measurement theories (CTT or IRT) provide several methodological models in language testing by specifying the relationships between measures, or observed scores and factors that affect these scores. CTT provides different ways in which we can estimate reliability. G-theory as an extension of CTT overcomes many of its limitations in that G-theory enables test developers to examine several sources of variance simultaneously, and to distinguish systematic from random test error. IRT presents a more powerful approach in that it can provide sample-free estimates of individual's true scores, ability levels, and the associated measurement error at each level.

These measurement theories, particularly the CTT model, are also very useful for the estimation of reliability in a language test. However, the reliability estimates based on the classical test theories are inappropriate for use with criterion-referenced test (CRT) due to the differences in the types of comparisons and decisions made. A CRT refers to a test that measures a student's performance according to a particular standard or criterion which has been agreed upon. The student must reach a certain level of performance to pass the test, and his score is therefore interpreted with reference to the criterion score rather than to the scores of other students in a norm-referenced test (NRT). In a CRT, reliability is concerned with both the dependability of test scores as indicators of an individual's level of mastery in a given domain of abilities and also the dependability of decisions that are based on the test scores.

Another limitation of CTT and IRT lies in their application in performance-based language tests, particularly those CRT performance tests that are designed for placement and diagnostic purposes. Recent studies on the communicative nature of language ability have brought back the interests in performance assessment among language test developers. Performance assessment is a test of authentic tasks that require examinees to demonstrate certain abilities. This type of assessment has been commonly used in language testing. As Norris et al. (1998) pointed out, “virtually all language tests have some degree of performance included (p. 7)”. Scholars have suggested that language tests should be viewed as performance oriented along a continuum of authenticity. The increasing attention language performance assessment receives is accompanied by criticism and concerns with regard to its reliability and validity (e.g. McNamara, 1997; Brown & Hudson, 1998). As Brown and Hudson (1998) pointed out, language performance assessment needs to satisfy the same standards as other types of language assessment. Within the previous CTT and IRT studies, however, it is difficult to obtain accurate

estimates of reliability and validity in a language performance test as other factors rather than examinee's ability may affect the test score. In other words, despite the fact that these measurement theories provide methodological tools for ad hoc and post hoc data analysis, the application of these two models may be limited if language testers do not have a thorough understanding of certain factors other than examinee's ability that may influence test score.

1.2 Factors that Affect Test Score

Language performance assessments, through the real-world assessment context, have introduced several factors that may influence examinee's test score. For example, the variability of rater judgments is considered a major source of measurement error in performance-based language assessment (Shohamy, 1983; 1984; Pollitt and Hutchinson, 1987; Lynch and McNamara, 1998). In previous studies, attention has been focused on four areas (Bachman, 2000):

a. Characteristics of Test Taker

Language testers examined different populations of test takers and found some common characteristics that may affect their test performance. Bachman (2000) summarized characteristics that have been studied, including test taker's occupation (Hill, 1993), aptitude (Sasaki, 1996; Sparks et al, 1998), background knowledge of test topic (Clapham, 1993; 1996) and personality characteristics (Berry, 1993).

b. Strategies of Test Taker

Examinees test taking strategies can be defined as certain test taking processes that the examinees are conscious of or have purposely selected (Bachman, 1998). Canale and Swain (1980) claimed that learners' ability to use language strategies constituents their strategic

competence. The finding of test taking strategies in oral and reading assessment supports Bachman's "interactional model" as the strategic competence represents how the components of language competence interact with each other.

c. Characteristics of Assessment Procedure

The interactive nature of language ability can also be represented by the effect of assessment characteristics on examinee performance. Significant relationship was found between item difficulty and the characteristics of test items (e.g. Anderson et al., 1991; Perkins & Brutten, 1993; Perkins, Gupta, L. & Tammana, 1995; Fortus, Corriat, & Fund, 1998). Other studies also found that different task types may generate different levels of test performance (e.g. Riley and Lee, 1996; Shohamy, 1994; Fulcher, 1996; McNamara & Lumley, 1997).

d. Rater Behaviors

Since language tests are more or less performance-oriented (Norris et al, 1998), the impact of raters' decision making becomes a recent focus in performance assessment.

One of the major preoccupations in the study of rater effect is the investigation of rater's decision making process, particularly in writing assessment. Scholars explored essay raters' decision making in holistic and other types of analytic scoring schemes in the context of English as a first and second language (Huot, 1990; Cumming, 1997; Hampy-Lyons & Kroll, 1997; Cumming, Kantor & Powers, 2001). More recent efforts have also been made in the rating process in the context of ESL assessment (Cumming, 1990; Vaughan, 1991; Shohamy, Gordon & Kramer, 1992; Weigle, 1994; Lumley, 2000).

The score of a language test represents a complexity of multiple influences. A language test score by itself is not necessarily a valid indicator of the particular language ability to be measured in a given test. The interactional nature of language ability determines that it is also

affected by the characteristics and content of the test, raters' characteristics and their scoring process, the characteristics of the test taker, and the strategies examinees employ in attempting to complete the test task. What makes the interpretation of test scores particularly difficult is that these factors undoubtedly interact with each other. This understanding of interactions in language testing suggests that careful considerations on different factors of a language test should be taken into account during the interpretation and use of test scores. Hence, in the context of writing performance assessment, the present study examines the effect of essay rater on test score, focusing on raters' scoring process and their decision making.

1.3 Rater Effects on Reliability and Validity

Reliability and validity are viewed as two distinct but related characteristics of test scores. It is agreed among language testers that reliability is a necessary condition to validity. In language performance test that requires raters, the distinction between these two characteristics can be quite blurred since rater variability may have a great impact on both test reliability and validity.

It is widely accepted that an important aspect of validity and reliability is concerned with the way raters arrive at their decisions (Huot, 1990). Therefore, it is fair to conclude that rater's decision making process is among the most important factors in the current trend of "interactive" or "communicative" language testing. This realization puts forward the demand on the development and facilitation of methodological tools to quantify rater's decision making process and also the interaction between rater and other stakeholders in a language test, hence providing a comprehensive interpretation of test scores.

1.3.1 Rater Effects on Test Reliability

All three major measurement theories have been applied as an attempt to interpret rater variation and rater reliability in performance test. In the traditional CTT model, the rater-related reliability is examined from a norm-referenced testing perspective, which is exemplified by rater consistency reliability. If rater variance is the major source of error in a given test, two reliability coefficients can be estimated based on rater consistence: the intra-rater reliability and inter-rater reliability. The former represents the consistency of the rating of an individual rater across different examinees, while the latter indicates the scoring agreement between two raters on the same examinees.

If a test involves more than one major random facet, for example, both tasks and raters are major sources of score variability, a multi-faceted analysis tool is required. G-theory can be used in such a context to analyze simultaneously more than one measurement facets. A number of studies have employed G-theory to examine the impact of rater variability on the dependability of test scores. Lynch and McNamara (1998) studied the rater and task variabilities as facets that contribute measurement errors to a performance-based assessment. Results from the G-study suggested that comparing to test task, rater is a more significant source of score variance.

In addition to CTT, Rasch model is another psychometric tool that is commonly used in examining the rater behavior in performance-based language assessment. Multifacet Rasch model provides the capability of modeling additional facets, hence making it particularly useful for analysis of subjectively rated performance tasks such as writing assessments. Weigle (1998) investigated the impact of rater training on their scoring by using the FACETS Rasch model. Rater behaviors before and after training were modeled using FACETS, which provides a four-

faceted IRT model with facets of examinee, writing prompt, rater and scoring scale. Results in this study indicate that raters' scoring experience has a significant effect on the severity and consistency of their scoring.

The application of mutlifacet Rasch measurement in rater differences and rater errors has also provided useful findings in test development and score interpretation. Gyagenda and Engelhard (1998) found a strong rater effect in writing assessment. The significant difference between essay raters indicates that for individual test taker it does matter who rates their essay as some raters are consistently more severe than others. This conclusion about persistent rater effect was also supported by other studies in writing assessment (Du & Wright, 1997; Engelhard, 1994). In addition to rater severity, other rater errors were examined in the study of Engelhard (1994). Significant rater differences were found in halo effect and central tendency, indicating that test rating is affected not only by test takers' performance but also by multiple rater factors.

1.3.2 Rater Effect on Test Validity

The pursuit of test validity remains an essential consideration for researchers and specialists in language testing. Messick (1989) illustrated his unified and faceted validity framework in a fourfold table shown in Figure 1.1. His theory cements the consensus that construct validity is the one unifying conception of validity and extends the boundaries of validity beyond the meaning of test score to include relevance and utility, value implications and social consequences. In other words, test validity refers to the degree to which the test actually measures the construct that it claims to measure, and also stands for the extent to which inferences, conclusions, and decisions made on the basis of test scores are appropriate and meaningful.

	Test Interpretation	*Test Use*
Evidential basis	Construct validity	Construct validity + relevance/utility
Consequential basis	Value implications	Social consequences

Figure 1.1: Messick's Framework of Validity.
Note: Adapted from "Validity," by S. Messick, 1989, *Educational Measurement*, NewYork: Macmillan.

While Messick's unitary conceptualization of validity was widely endorsed, many disagreed with his view of validity and found that his framework does not help in the practical validation process. Kane (2008) discussed the benefits and shortcomings of Messick's validity model and pointed out that "this unitary framework may be more useful for thinking about fundamental issues in validity theory than it is for planning a validation effect" (p. 77). His claims are consistent with findings of a recent study conducted by Cizek, Rosenberg, and Koons (2008). They reviewed 283 tests and found only 2.5 percent of these test had a unitary conceptualization of validity and few of them reported validity evidence based on consequences. In addition, only one quarter of the tests reviewed referred to test validity as a characteristic of test score, inference, or interpretation.

In late 1980s, Cronbach (1988) proposed that evaluation argument should be used in the validation of score interpretations and uses. He suggested that a validity argument helps generate a coherent analysis of all of the evidence for the proposed interpretation, thus providing an overall evaluation of the intended score interpretations and uses. Based on Cronbach's framework of validity argument, Kane further developed the concept of an argument-based approach to validity. He argued that validation should always begin with an interpretive argument that specifies a specification of the proposed interpretations and uses of the scores, and the validity argument then provides an evaluation of the interpretive argument. This approach has

been well received by developers and users of second language assessments. For example, a set of validity argument have been developed for the TOEFL iBT. Chapelle, Enright, and Jamieson (2010) endorsed Kane's framework of interpretive argument and argued that his approach provides conceptual tools to express the multifaceted meaning of test scores.

Within Kane' validity framework, an interpretive argument is articulated through a validation process that considers the reasoning from the test score to the proposed interpretations and the plausibility of the associated inferences and assumptions. Validators will then evaluate the inferences and assumptions by examining the validity argument developed from the interpretive argument, gathering different types of validity evidence to support the validity argument as claims, intended inferences, and assumptions. For a placement testing system, an interpretative argument includes four major inferences: scoring, generalization, extrapolation, and a decision. Each of the inferences depends on a set of assumptions that must be evaluated. Scoring, as the first inference in the interpretive argument, employs a scoring rubric as a guideline for student performance to assign a score to each student's performance on the test tasks. This process makes inference from observed performance to observed score. The scoring inference relies on two assumptions, 1) the scoring rubric is appropriate, and 2) the scoring rubric is applied accurately and consistently by rater. The degree of confidence about scoring inference provides information about the quality of the examinee's responses. As evidence, rater's scoring procedures, judgments of examinee's responses, and scoring methods in test specifications should be gathered and analyzed as important measures of score precision.

As test raters are deeply involved in the interpretative argument for performance testing, an important aspect of validity argument is associated with how the process of rating is managed (Lumley, 2002). Rating related factors are fundamental to the traditional direct writing

assessment as depicted in Figure 1.2, which provides a summary of the shared procedures in most writing assessments, the purpose of these procedures and the assumptions upon which they are based.

Procedure	*Purpose*	*Assumption*
Scoring Guideline	Recognize features of writing quality	Writing quality can be defined and determined
Rater Training	Foster agreement on independent rater scores	One set of features of student writing for which raters should agree
Scores On Papers	Fix degree of writing quality for comparing writing ability and making decisions on that ability	Student ability to write can be coded and communicated numerically
Interrater Reliability	Calculate the degree of agreement between independent raters	Consistency and standardization to be maintained across time and location
Validity	Determine the assessment measures what it purports to measure	An assessment's value is limited to distinct goals and properties in the instrument itself

Figure 1.2: Direct Writing Assessment: Procedures, Purposes and Assumption
Notes: Note: Adapted from "Toward a New Theory of Writing Assessment," by B. Huot, 1996, College Composition and Communication, Vol. 47, No.4. p. 551.

From Figure 1.2, we can see that the preparation and the production of rating account for most factors in test procedure. Though this may sound evident, an dependable rating process is in fact a prerequisite of test validity for writing performance tests. That is to say, a writing test is not able to measure the targeted writing ability unless raters actually comprehend the writing responses and evaluate the essays based on the required scoring schemes. Otherwise, the test score fails to represent or represents less precisely test takers' ability level for the target construct, even though other factors, such as test content, response process, the internal structure of the test and the consequences of testing, are perfectly controlled. For example, an integrated

ESL writing test is designed to elicit college students' ESL academic writing ability. The grade represents test taker's ability and can be compared to related non-test situations if and only if essay grading is based on raters' comprehension of text content and their accurate interpretation of scoring criteria in language related terms. Otherwise, essay scores may reflect construct-irrelevant variability, such as the neatness of handwriting or the writers' creativity. As a result, the test administrators would not be able to make accurate inferences from or interpretation of the test score, failing to make any appropriate decisions or conclusions based on the inferences from performance.

As composition grading is necessarily based on raters' subjective judgment, the way that raters comprehend writing responses and arrive at their decisions has a great influence on the validity of writing assessment. Researchers have addressed their interests in raters' decision making by 1) investigating in various factors that may affect raters' decision (Huot, 1990; Cumming, Kantor and Powers, 2002); and 2) indirectly studying raters' decision making process by looking at the final score productions. Nevertheless, the effect of essay rater as the executor of rating process and user of rater schemes still remain underrepresented in the study of test validity. Very little information has been obtained on what effects raters' essay reading and their rating process have on the achievement of test validity.

1.3.3 Limitation of Measurement Theories in Rating Study

In order to examine how raters affect the reliability and validity in a performance assessment, the essential question is how raters arrive at their scoring decision when grating examinees' responses. Currently used measurement approaches are essentially silent on this point. As Hambleton, Swaminathan, and Rogers (1991) noted that "much of the IRT research to

date has emphasized the use of mathematical models that provide little in the way of psychological interpretations of examinee item and test performance" (p. 164). Cumming (1990) also pointed out that, particularly in writing assessment, "direct validation of the judgment processes used in these assessment methods has not been possible because there is insufficient knowledge about the decision making or criteria which raters or teachers actually use to perform such evaluations" (p.32).

Within the framework of CTT and IRT, most researches analyze rater's decision making process by looking into the scoring scheme and the scores assigned by rater. For example, Congdon and McQueen (2002) investigated the stability of rater severity on the writing performance of elementary school students by examining rater's scoring data over an extended rating period. Stuhlmann and her colleagues (1999) explored the training effect on rater agreement and consistency in portfolio assessment by quantifying the pre-training and post-training essay scores assigned by both experienced and inexperienced raters. Shohamy, Gordon and Kramer (1992) also collected test scores from raters with different background to examine the influence of training and raters' background on the reliability of direct writing assessment.

Unfortunately, this indirect approach could not be able to keep track of the "online" record of rating process. Very little if any attention has been paid directly on the very process of rater's decision making. So based on what criteria does a rater assign a score to a written composition? Why does a rater choose a particular score from the rating scales? If raters assign different scores to the same essay, what is the source of the disagreement? Is it because raters have different expectations, and different backgrounds or because they actually went through a totally different decision making process? Most of these questions still remain unanswered.

Another important criticism about the application of measurement theories is addressed

on their assumptions. Despite the fact that CTT and IRT have been widely used in language testing, these two models were originally designed for psychological measurement. Their basic assumptions are inconsistent with the widely accepted understanding of language proficiency in the field of applied linguistics. As theories of measurement in general, CTT and IRT assume that there is one measurement construct. In the context of language test, for example, this construct per se can be roughly defined as a narrow conception of "language proficiency", which is an isolated "trait". CTT and IRT share a common assumption about the unitary feature of this construct to be measured: CTT assumes there is a "true score" of an individual's ability and G-theory as part of the CTT model employs the basic idea that there is a universe score which is the analog of CTT's true score; most of the IRT models currently used in language testing hold the unidimensionality assumption, indicating that there is a unique trait which roughly corresponds to the language ability of the test taker.

In language testing, however, the target construct –language proficiency or communicative language ability refined by Bachman (1990)—is thought to be a multi-componential ability. Built upon Canale and Swain's four-component description, Bachman's communicative competence, or "organizational competence" can be divided into grammatical and discourse (or textual) competence and pragmatic competence (1990). The multiconponential nature of language proficiency determines that examinee's communicative competence does not always develop at the same rate in all domains. Therefore, models that posit a single continuum of proficiency are theoretically limited (Perkins & Gass, 1996).

Such a discrepancy between the definition of test construct in measurement models and that in language testing may raise problems in test validity. The current trend of communicative approach and the corresponding performance assessment attempt to measure test taker's

communicative language ability, which consists in a comprehensive evaluation of the different components of test taker's communicative competence. The shift of the focus of language testing from formal language to communicative language ability comes under the criticism about test validity. According to Messick (1989), test validity is an "integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of interferences and actions based on test scores" (p.13). Within the current framework of communicative approach, the inferences from test score are particularly useful not only in language teaching and learning, but also in the research of language learner's developmental sequence. A general statistic in terms of overall language proficiency, however, does not provide useful information in this sense, thus jeopardizing the overall test validity.

Different understanding of measurement error is another concern in the application of current measurement models in language testing. In the true score approach, measurement error is defined as the deviation of test score from the "true" score. In language performance tests, however, this definition of error does not fit in the "interactive" framework in which there is a significant amount of interaction between test taker, test task and rater (Bachman, 1990; 2000). The effort of G-theory in discerning the source of errors and measure the scale of variance introduced by difference sources (including rater and task type) is also limited as it is not able to further explore the structure and magnitude of these interactions. Hence, whether certain variances are pure measurement errors or whether they are associated with a specific interactive pattern is unknown in the true-score framework.

In the performance test that requires rater, the problem associated with the error definition also exists. Linacre (1989) noted that in true-score approaches, rater variation is considered as undesirable error variance, which must be minimized to make the test reliable.

This understanding of rater variation, however, has practical and theoretical problems. First of all, the absolute agreement between raters never happens in the real world test practice. Even though raters could be trained to have a total consensus on the score assigned to the same examinee, questions about the interpretability of test scores would still remain since the rating scale may not be linear (Weigle, 1998). The many-faceted Rasch model takes a different approach to the phenomenon of rater variation. In this approach, rater variation is seen as an inevitable part of the rating process. Rather than a hindrance to measurement, rater variation is considered beneficial as it provides enough variability to allow probabilistic estimation of rater severity, task difficulty, and examinee ability on the same linear scale (Weigle 1998).

This discrepancy causes confusion in understanding the purpose of rater training in performance tests. In the literature of measurement, the purpose of rater training is primarily associated with the feasibility of increasing reliability in ratings. However, researchers have not reached a consensus on if an effective training should enhance rater agreement or not. The function of rater training has been addressed from different perspectives. Researchers argued both for and against emphasizing agreement in rater training in according to different measurement approaches they are taking (Barritt, Stock & Clark, 1986; Charney, 1984; Lunz, Wright, & Linacre, 1990; also see Weigle, 1998).

Again, this confusion is rooted in the lack of the understanding of rater's decision making process. The surface disagreement or agreement does not provide enough information about how raters reach their score assignments. For example, the score of 4 assigned by one rater does not necessarily mean the same as a score of 4 assigned by another rater. These two raters agree with each other on this examinee's performance only when these two scores are assigned through the same decision making process. Without the knowledge of this rating process, it is impossible for

test practitioners to decide whether rater disagreement should be reduced. As neither CTT nor IRT has directly tapped into the rating process, the error definition in these models, particularly with regard to rater, is of concerns in language testing.

Last but not least, the basic assumption on the characteristics of a target construct is different in psychological measurement and language testing. As a psychometric approach, IRT is a latent variable analysis which deals with variable that are not directly observed. Without any measurement error, a latent variable is also known as a hypothetical construct, whose existence is to be measured by multiple indicators. In language assessments, however, the target construct is well defined and observable. For example, in a direct writing assessment, the target language proficiency can be defined as examinee's communicative writing ability within a certain situation. Rather than measure this writing ability through other language indicators such as grammar and vocabulary, the target construct can be measured directly in a performance test which reflects tasks that an examinee may have to perform in the real world. Language test, comparing to psychological measurement, is a totally different type of measurement because its target construct is observable and measurable. Therefore, the application of latent variable models in the study of language performance testing has both theoretical and empirical limitations.

In conclusion, the implementation of measurement theories in language testing has been consistently challenged during the theoretical advances in this field. With the development of these performance-based language tests, language testers have been faced with complex problems that have both theoretical and practical implications. One of these problems is that language testers do not have enough understanding of different factors that affect test score, thus failing to avoid bias for test development and for score interpretation (Bachman, 1990). Another

problem, as Bachman pointed out, is “determining how scores from language test behave as quantifications of performance” (p. 8). In order to solve these two problems within the communicative approach of language testing, a comprehensive investigation of the rating process would be of great necessity.

1.4 Rater Effect in Writing Assessment

1.4.1 Scoring Procedures for Writing Assessment

Different types of scoring schemes and their construct validity for essay scoring have been evaluated for their effect on essay scoring, both in the contexts of English as the first language (Charney, 1984; Huot, 1990; Purves, 1992) and English as a Second or Foreign Language (ESL/EFL; Brindley, 1998; Connor-Linton, 1995; Cumming, 1997; Hamp-Lyons & Kroll, 1997; Raimes, 1990). In the literature of writing assessment, three major rating criteria have been developed to evaluate student's writing, including the Primary Trait scoring, holistic scoring and analytic scoring (Weigle, 2002).

Primary Trait scoring is best known as the rating criteria used in the National Assessment of Educational Progress (NAEP). The rating scale in Primary Trait rubrics consists of: (1) a specific writing task, (2) a statement of the primary rhetorical trait, (3) a hypothesis about the expected performance on the given task, (4) a statement of the relationship between the task and the primary trait, (5) a rating scale which represents each performance level, (6)sample scripts at each score level, and (7)explanation of the sample script scored at a certain level (Weigle, 2002). The Primary Trait scoring criteria is task sensitive and requires raters to understand examinees’ writing performance within a well-defined discourse range. Therefore, it is most frequently

applied in a school context. Though it may provide diagnostic information about students' writing abilities, Primary Trait assessment hasn't been widely used in ESL writing test.

First developed by Diederich (1974), analytic scoring involves specific aspects of a writing sample in various components. This scoring procedure focuses on several identifiable features of a good writing, such as essay organization, development, vocabulary, grammar and other essay qualities. In Diederich's framework of analytic scoring, raters give scores to individual identifiable traits and these scores are tallied or sometimes weighted to provide rating for an essay. This scoring scheme has been suggested as the most reliable of all direct writing assessment procedures (Scherer, 1985; Veal & Hudson, 1983; also cited by Huot, 1990). Compared to the holistic procedure, analytic scoring provides more diagnostic feedbacks to guide instruction. Therefore, it is more helpful for ESL learners who tend to show different performance across different scoring aspects/dimensions (Hamp-Lyons, 1995, 1991; Weigle, 2002). A major disadvantage of this scoring scheme is that it takes more time than holistic scoring, which limits its application in large scale assessment due to the large scoring expense (Weigle, 2002; Lee, Gentile and Kantor, 2005). In addition, as previous studies have shown that holistic scores correlate reasonable well with those generated by analytic scoring (Freedman, 1984; Veal & Hudson, 1983), holistic scoring is usually more recommended, especially for large-scale writing tests.

As the most commonly used scoring scheme in ESL writing assessment, holistic scoring reflects rater's general impression of the quality of a piece of writing. In most holistic rating procedures, scoring guidelines detail which general characteristics represent writing quality for each score of the scale being used. Although holistic scoring is generally not quite as reliable as analytic scoring, it correlates well enough to be a viable alternative (Baue,. 1981; Veal &

Hudson, 1983). White (1985) also pointed out that holistic scoring is more valid than analytic scoring because the rating process represents a more authentic reaction a reader has to a written passage; while a analytic scoring requires raters to focus on the writing components instead of looking at the overall meaning of a passage (also cited in Weigle, 2002). From a practical point of view, holistic scoring is faster and less expensive (Weigle, 2002). At any rate, holistic scoring has been viewed as the most economical of all direct writing procedures (Bauer, 1981: Scherer, 1985: Veal & Hudson, 1983) and therefore the most popular (Faigley, et al., 1985: White, 1985). Decisions about which evaluation procedures should be selected need to be made within the context of a specific testing situation (Huot, 1997). In the current study, holistic scoring schemes are used to evaluate the essay quality in the EPT writing test at the University of Illinois at Urbana-Champaign (UIUC).

1.4.2 Factors that Affect Essay Rater's Judgment

The literature of writing assessment has shown that some categories of writing responses have greater impact on essay rater's scoring judgment. Though studies on these factors may not be able to directly capture rater's decision making process, it still provides valuable insights about based on what criteria raters arrive at their scoring decision.

a. Essay Features

The relationship of textual features and essay scores has interested researchers for many years. The earlier studies focused on syntax and various indexes, whereas the later works were more interested in global-level language features. This shift in the type of textual analysis is obviously related to the shift in linguistic theory. With earlier studies having a link to Chomsky's generative grammar, the later interest in global-level textual examination has been fostered by

the developments in linguistics, especially in intersentential grammars like Cohesion and Functional Sentence Perspective.

In the early study of text features, the T-unit (an independent clause) used to be the major form of textual analysis, and it was used to determine syntactic maturity and, therefore, writing quality (Hunt, 1965; O'Donnell, Griffin & Norris, 1967). The results of these early studies indicate that T-units appear to be most sensitive to the writing of elementary school children, an age at which syntactic development is still occurring. Veal (1974) found a strong correlation between T-unit length and quality in the writing of 2nd, 4th, and 6th graders. Stewart and Grobe (1979) also found a relationship between T-units and writing quality in 5th graders' writing, which was not evident in the writing of 8th and 1lth graders. These findings were supported by Witte et al. (1986), who discovered that raters were most influenced by writings that exhibited the lowest levels of syntactic complexity. Other studies that have attempted to determine the effects of syntax in the writing of high school and college students have been unable to find any correlations between syntax and writing quality (Crowhurst, 1980; Greenberg, 1981; Grobe, 1981; Nielsen & Pichi., 1981; Nold & Freedman. 1977; Stewart & Grobe, 1979). It seems that the studies that examined writing of lower-level syntactic complexity tend to identify a relationship between syntax and writing quality.

Previous research has also examined the effect of syntactic accuracy on the evaluation of essay quality. Li (2000) investigated the relationship between computerized scoring and human scoring of ESL writing samples using measures of syntactic complexity, lexical complexity, and grammatical accuracy. The author found that the only statistically significant correlations that were observed between computer and human scoring were between both computerized measures of grammatical accuracy and the human-evaluated measure of grammar. Based on prior literature

on natural language processing, Educational Testing Service (ETS) has developed an e-rater to score TOEFL writing samples by evaluating nine writing features and two content features. The nine writing features include five error features of grammar, such as agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical errors (Attali & Burstein, 2005; Ramineni, et. al., 2012).

Another important factor that influences essay rater's judgment is word choice. Grobe (1981) found that what raters perceive as "good" writing is closely associated with vocabulary diversity. Neilsen and Pichi (1981) also reported that lexical features have a significant impact on rater judgment. They did not find a significant relationship, however, between syntactic complexity and rater perception of writing quality. Chinn (1979) reported on two studies that link vocabulary development to effective elementary-level language pedagogy and the success on a high school writing competency examination. A lexical analysis revealed a direct correlation between competency rating and effective verb use. Chinn concluded that verb choice is a significant predictor of writing quality as assessed through holistic scoring.

Research has shown that rapid or automatic decoding are strong predictors of text readability. Previous studies suggest that high proficiency writers tend to use less frequent words in writing (Just & Carpenter, 1987; McNamara, Crossley, and McCarthy, 2010). A more recent study conducted by McNamara, Crossley, and McCarthy (2010) used an automated tool to examine a corpus of expert-graded essays, based on a standardized scoring rubric, to distinguish the differences between the essays that were rated as high and those rated as low. They found that word frequency is one of the three most predictive indices of essay quality.

Other studies have looked at writing quality by investigating the relationship between essay quality and text length (e.g. Homburg, 1984). Chodorow and Burstein (2004) studied the

accuracy of two versions of e-rater, when the effect of essay length was removed from one of them. They used both e-raters to rate thousands of essays written for the computer-based version (CBT) of the TOEFL on seven prompts. They found that scores produced using length as the only predictor matched holistic scores half of the time and came within one point of holistic scores 95% of the time. Similar results were also found in a more recent study that explored the use of objective measures to assess writing quality (Kyle, 2011). In this study, Coh-Metrix 2.0, an online text analysis tool, was used to measure 54 linguistic properties of argumentative essays written by ESL students and English as a Foreign Language (EFL) students. Using discriminant function analysis, Kyle reported that essay length was able to significantly discriminate between holistically evaluated high and low quality essays. He found that high quality essays tend to be longer, with an average length of 642.21 words; while low quality essays have an average length of 495.42 words. This study also found that overall sentence length and word length are also strong predictors of essay quality. Overall, EFL essays tend to be perceived by human raters of higher quality if they use longer sentences with longer words. In addition, studies that examined how linguistic features can predict essay scores in integrated writing tasks have shown that essays that contain more words are more likely to receive higher scores (Cumming, et al., 2006; Watanabe, 2001).

Another approach of textual analysis focuses on the application of intersentential grammars that attempt to explain how meaning is projected across the entire writing. The attempt to gauge the impact of textual features beyond immediate sentence boundaries is a reflection of new developments in linguistics that are concerned with global-level textual features. One important research interest is the cohesion of a composition (Bamberg, 1983; Fahenstock, 1983; Witte & Faigley, 1981). Cohesion in English depicts a systematic use and taxonomy of cohesive

ties that "accounts for the essential semantic relations whereby any passage of speech or writing is enabled to function as a text" (Haliday & Hasan, 1976, p. 13). This interest in cohesion has evolved into a series of research studies about the relationship of cohesion and essay quality. However, contradictory results were found from different researchers. Witte and Faigley (1981) claimed that high-quality writing had a greater cohesive density (rate of cohesive ties) than did low-quality writing. Tierney and Mosenthal (1983) analyzed 24 essays written by high school seniors for cohesion and had the same essays rated for coherence. They found no relationship between cohesive density and coherence. Their results, however, was challenged by McCulley (1985). Although he found no correlation between cohesive density and writing quality, McCulley's finding did contradict the results from Tierney and Mosenthal (1983) by indicating that "the evidence presented in this study strongly suggests that textual cohesion is a sub-element of coherence." Neuner (1987) analyzed 40 high- and low-quality essays. Although he concurred with earlier findings about cohesive density not being a predictor of writing quality, he did suggest that chains of cohesive ties can be used to distinguish writing quality in student writing. Zhang (2000) investigated the relative importance of various grammatical and discourse features in the evaluation of second language writing samples and found that raters considered cohesion as an important element in judging essay quality. Crossley and McNamara (2010) also argued that coherence is an important attribute of overall essay quality, but that expert raters evaluate coherence based on the absence of cohesive cues in the essays rather than their presence.

It seems that there is no consensus on whether coherence or cohesion plays important roles in judgments of essay quality. However, empirical studies have shown that cohesion or coherence facilitates text comprehension (McNamara, Louwerse, McCarthy, & Graesser, 2010). Research found that that increasing the cohesion of a text significantly facilitates and improves

text comprehension for both skilled and less-skilled readers (Gernsbacher, 1990; Beck et al., 1984; Cataldo & Oakhill, 2000; Linderholm et al., 2000; Loxterman et al., 1994).

The findings of recent studies have clearly indicated that the interest of textual analysis and essay quality have been placed in the discourse-level research. In addition to the attending to essay cohesion (McCulley, 1985; Neuner. 1987; Tierney & Mosenthal, 1983), more investigations have been conducted with topical structure (Witte, 1983a, 1983b) and information in noun phrases (Sullivan. 1987). Although this work is still in its formative stages, it is evident that there are an increasing number of discourse-level studies exploring the reading and rating of student writing.

b. Raters' Response Categories

Diederich et al. (1961) analyzed over 11,000 scoring comments, responses and annotations made by essay raters for college freshmen. By using factor analysis to interpret the correlations between raters, Diederich and his colleagues were able to isolate five main types of rater responses including: 1) Ideas and their relevance, clarity, development and persuasiveness; 2) Form and its organization and analysis; 3) Flavor, including style interest and sincerity; 4) Mechanics such as grammar and punctuation errors; and 5) Wording, which stands for the selection and arrangements of words (Diederich et al., 1961). The validity of these five categories of responses was tested by Jones (1978), who reported that these categories represent all comments made by his raters. This conclusion indicates that the five categories are an accurate description of rater response to student writing.

Based on Diederich et al's framework, studies by Freedman (1979, 1981, 1984; Freedman & Calfee, 1983) represent some of the most informative research conducted on the influences of student writing on raters. Freedman (1979) rewrote students' essays to make them

either strong or weak in the categories of content, organization, sentence structure, and mechanics. Analyses of variance showed that raters were most affected by content and organization: content was proved to be the most significant feature, followed by organization. Mechanics and sentence structure ranked third and fourth, respectively. Freedman concluded that holistic raters base their judgments primarily on the content and organization of student writing. It is important to note that mechanics and sentence structure were only important influences when organization was strong.

The importance of various response criteria was further examined by Breland and Jones (1984), who correlated raters' holistic scores with comments made on the same papers. As an attempt to identify the criteria that raters use to make judgments when rating holistically, their study suggested that organization, support, and ideas were the three most important considerations in rater judgment of essay quality. This finding was confirmed by rater's response of a poll about what characteristics raters perceived as important in student writing before starting the rating session. The researchers found the results of the poll were consistent with the ratings given to essays during the scoring session. Breland and Jones thus concluded that raters are not only affected by certain criteria when grading holistically, but also aware of the criteria on which they base their judgments of writing quality.

At this point, it appears that some contradictory results have been observed from the previous studies about the impact of writing responses on rater's decision making. Despite the effort of researchers in writing assessment, little consensus has been reached with regard to *what* and *why* particular scoring criteria have the most impact on rater's judgment. By far, the notion of whether or not raters score essays the way they think they do or the way they are expected to do has not been fully explored in the literature. Most studies measure only raters' responses to

manipulated categories rather than capture the very process of how raters comprehend the text and how they arrive at their scoring responses. In other words, what is missing in the picture is the “online” evidence of rater’s decision making process. Another limitation of the previous studies is that rater’s role as a text reader is underrepresented. Though it is well acknowledged that rater’s scoring responses are affected by essay features or scoring categories, it is still not clear when this influence occurs during the reading comprehension and how this may affect rater’s judgment. Unfortunately, the methodology used in the literature only allows the researchers to look at raters' final judgments or compare their verbal comments. Most previous methods are not able to capture the time-by-location information of rater’s essay comprehension or make the time-by-location comparison of rater’s scoring responses, though these variables may contribute to a more complete picture of rater’s decision making process.

In direct writing assessment, the rater variability affected by these above-mentioned variables is inevitable because it is “part of the natural process of reading" (Stock & Robinson.1987. p. 105). Therefore, a consideration of the way raters read would be necessary to reveal some important but often neglected connections between phenomena associated with essay rating and the reading process.

1.5 Rater’s Reading Comprehension during Essay Grading

1.5.1 Understanding the Rating Process: Indirect Approaches

As stated above, it would be of great help for language testers if we understand better rater’s decision making and rater’s influence on the validity of test scores in a performance-based language assessment. Previous studies employing classical measurement approaches tapped this

issue indirectly by looking at the possible factors that may affect rater's decision making rather than the process per se. As an attempt to directly examine the rating process, many recent studies have follow the method of think-aloud protocol described by Ericsson and Simon (1993), which requires raters to describe the rating process in verbal reports as they assign the grade (Cumming, 1990; Vaughan, 1991; Weigle, 1994; Lumley, 2000; Cumming, et al., 2001).

Vaughan (1987) collected the talk-aloud protocols of nine experienced raters scoring six compositions according to the 6-point CUNY scoring rubric. Results indicated that content received the most comments, but handwriting was second. Great variation in rating strategy was also found among raters. Huot (1988) recorded the think-aloud protocols of eight raters reading 42 student essays, but found no difference in rating criteria between the two rater groups. Content and organization received the most attention from two groups in the study. Cumming et al. (2001) adopted the think-aloud method in examining essay raters' decision making behavior and factors that affect their scoring decision. In this study, a comprehensive list of 35 decision making behaviors was collected from experienced raters as the decision making framework. Their findings suggest that raters focus on different scoring criteria when they grade essays of different quality or essays written by L1 student or ESL student.

This think-aloud approach has its own limitation as well. First of all, the relationship between scale content and text quality still remains obscure in this approach. The behavioral data from rater's oral report is subjective, difficult to process and almost impossible to quantify. Researchers have also claimed that this approach addresses the artificial scoring process as the think-aloud behavior may interfere with rater's decision making process.

To sum up, though numerous studies have been conducted on the rating process of writing performance test, there is no consensus about how direct evaluation procedures affect

rater's ability to judge writing quality. There still remain many unanswered questions with regard to how raters reach their final judgment or how essay quality affects rater's perception as a reader. As Huot (1990) proposed, more research about the influences of essay categories on rater judgment would be necessary and these studies "should focus on the raters themselves, the nature of the fluent reading process, and the process of reading according to specific guidelines, especially for the purposes of agreement".

1.5.2 Essay Rater's Reading Comprehension

In order to evaluate rater's scoring judgment, first we need understand what contribute to their decision making process. The rating process can be divided into two major stages: 1) text reading and comprehension, and 2) scoring. Though raters assign a score after text reading, their decision making, however, is based on the interaction of these two rating procedures. Therefore, raters' reading comprehension and scoring are inseparable components of their decision making process. A consideration of the way raters read can help us to understand some important but often neglected connections between the phenomena associated with essay rating and the process of raters' reading comprehension of students' essays.

Reading is not a single-factor process. It is a multifaceted procedure which consists of behavioral variables including eye movement, word recognition, lexical and syntactic processing, meaning accessing and inference making. Comprehension comes into the stage of processing when word recognition and parsing are finished. As a result of identifying words and parsing sentences, readers need to identify their thematic roles and access their individual meanings. The next task for reader is to integrate these different aspects into sentence representation, to integrate it with what have gone before, and to decide what to do with this

representation. Though reading and comprehension represent different stages of text processing, they are intertwined in nature. It is plausible to suggest that reading and comprehension are closely correlated as comprehension can be seen as a product of the coordination of various reading variables (Rayner, et al., 2006). In other words, study of reading variables can provide valuable information regarding moment to moment comprehension process ((Rayner, et al., 2006, Rayner, 1998).

When reading students' essays, as readers/raters are asked to read for meaning, their reading comprehension requires a multivariate skill involving a complex combination and integration of a variety of cognitive, linguistic, and nonlinguistic skills. These skills range from the very basic low-level processing abilities such as text decoding to high-level skills of syntax, semantics, and discourse, and even to the knowledge of text representation and the integration of ideas with the readers'/raters' global knowledge. There has been an ongoing debate in the reading research literature with regard to the relative importance of each of these processing levels in reading comprehension. However, for the reading comprehension of a long passage such as essay reading in a writing tests, many researchers have argued for the primacy of higher-level syntactic, semantic, and text integration processes, minimizing the role of basic lower-level word recognition processes in fluent reading (Goodman, 1971, 1996; Smith, 1971, 1994). Study of these higher level processes is also remarkably informative as to understanding raters' reading comprehension of an essay.

1.5.3 Reading Comprehension and Eye Movement

An important issue in reading concerns when and where readers move their eyes. As Staub and Rayner (2006) pointed out, "eye movement is the natural part of the reading process,

… the information about where readers fixate in the text and how long they look at different part of the text provides remarkably reliable data about comprehension at a number of levels". In this case, the pattern of readers' eye movement and its temporal representations – the reading rate and total reading time can be viewed as robust indicators of text comprehension. Previous studies about eye movement found that readers' eye fixation time is affected by 1) the properties of an individual word; 2) the syntactic anomaly of a sentence; and 3) the coherence of a discourse.

1.5.3.1 Eye movement at word level

One variable that affects readers' eye fixation is word length. Just and Carpenter (1998) first reported that readers' gaze duration becomes longer as word length increases (also see Rayner et al., 1996). This effect can be accounted for by the fact that as words get longer, the probability of readers' refixation on this word increases (Rayner, 1998).

Another variable that gets more attention in the study of reading is word frequency, which is determined by counting the occurrence of a word in a corpus of printed or spoken materials. Though it is often viewed as confounded with word length, word frequency has a strong influence on fixation time when word length is controlled. Many studies have found that readers look longer at low-frequency words than at high-frequency words (Altarriba, et al., 1996; Henderson & Ferreira, 1990, 1993; Inhoff & Rayner, 1986; Just & Carpenter, 1980; Raney & Rayner, 1995; Rayner, 1977). Rayner (1977) and Just and Carpenter (1980) reported that readers' gaze duration is longer when they look at low-frequency words. After controlling for word length, Rayner and Duffy (1986) and Inhoff and Rayner (1986) also found a significant frequency effect both on the first fixation on a word and on gaze duration. When reading high frequency words, however, readers tend to skip those words more often than low-frequency

words, especially when words are six letters or less (O'Regan, 1979; Rayner et al., 1996).

This frequency effect on word reading can be accounted for by Morrison (1984)'s eye movement model and its subsequent variations. Morrison first suggested that readers' attention shift and subsequent eye movement are triggered by the encoding of the fixated words. Henderson and Ferreira (1990) and Pollastek and Rayner (1990) proposed equated encoding of the fixated word with lexical access to that word. In their models, lexical access is the process by which a word's orthographic and/or phonological pattern is identified so that the semantic information can be retrieved. As lexical access is assumed to be influenced by word frequency, fixation time on low-frequency words may be longer than on high-frequency words.

1.5.3.2 Eye movement at sentence level

Syntactic anomaly of a sentence is another factor that affects reader's eye movement, thus it has been the focus of many scholars in text reading (Braze et al., 2002; Deutsch and Bentin, 2001; Ni et al., 1998; Pearlmutter et al., 1999). For example, Pearlmutter et al. (1999) had participants read sentences in two conditions: 1) the verb either did or did not agree with the subject in number; 2) an irrelevant noun that intervened between the subject and the verb could either agree with the verb or not. Pearlmutter and colleagues reported that reader's gaze duration increases when reading sentences in both conditions. Deutsch and Bentin (2001) found that the gender mismatch between subject and verb causes a first pass effect on the verb and the sum of all fixations on the verb is longer. Sturt (2003) also found that if an anaphor, such as *himself, herself* did not match the stereotypical gender of its antecedent, reader's first fixation on the anaphor has a longer duration time.

Another question about the relationship between syntactic processing and eye movements

is, in the absence of ambiguity, whether reading time is affected by syntactic complexity. For example, when a sentence has a longer sentence length or a larger number of nodes in the sentence's phrase structure diagram, the total reading time may vary accordingly. Though this topic has received relatively little investigation in English, some findings were reported on the reading study of European languages. For example, Hyönä and Vainio (2001) examined how morphologically complex clause constructions were processed during reading Finnish. Reader's eye fixation patterns were recorded when they read two alternative versions of the same linguistic construction, a morphologically complex converb construction and its less complex subclause counterpart. Results indicate that more complex converb constructions produce longer gaze durations than the subclause constructions that have the same length and frequency. However, the complexity effect is reversed when the more complex clause form is clearly more common in the language than its less complex counterpart. This finding suggests that both structural complexity and structural frequency influence the ease with which linguistic expressions are processed during reading.

1.5.3.3 Eye movement at Discourse Level

The comprehension of a text is a much more complex process comparing to word or sentence level comprehension. In addition to word recognition and syntactic parsing of a sentence, readers must also maintain a representation of the entities that have been mentioned and relate the information that is currently being processed to this stored representation. This process requires readers to determine, for example, what entities pronouns and definite descriptions refer to, and make inferences about relationships between events and entities (Staub and Rayner, 2006).

Compared to the large number of eye movement studies of syntactic parsing, relatively fewer studies have examined how such discourse processing affects eye movements in reading. Among these studies, the constructivist principle—*search after meaning*—has been adopted in discoursing processing. As one basic assumption of this principle, researchers believe that readers attempt to construct a meaning representation that is coherent at both local and global levels. Local coherence, or cohesion, refers to "structures and processes that organize elements, constituents, referents of adjacent clauses or short sequences of clauses" (Graesser, Singer, and Trabasso, 1994). Global coherence stands for the established organization and the interrelation between the local information and the higher-order discourse-level information. Previous investigations have demonstrated that an incoherent discourse is more difficult to process, thus increasing "the duration of eye fixations as well as the number of fixations and the probability of regression during silent reading of long passages of text" (Rayner, Chace, Slattery, and Ashby, 2006).

The inconsistency between an anaphor and the antecedent has been investigated as one of the major accounts for an incoherent passage. Generally speaking, an anaphoric element such as a pronoun or a reflexive typically has an antecedent. If the anaphor and related antecedent are mismatched, readers may have difficulty constructing the discourse coherence, thus slowing down their reading rate. For example, if the antecedent violates a gender stereotype, reading time on the pronoun is inflated (Duffy and Keir, 2004; Sturt, 2003; Sturt and Lombardo, 2005). Cook (2005) investigated the effect of anaphors and their antecedents if they are inconsistent but semantically high overlapping or low overlapping. Cook found a longer reading time on the region following the anaphor. The rereading time on the anaphor suggested processing difficulty in the inconsistent condition. These results suggest that readers noted the inconsistency and

attempted to resolve it by rereading the anaphor or by spending more time on the spillover region. This longer reading time can also be explained by the regression data, which indicates that more regressions out of the postanaphor region occurr in the inconsistent conditions. In addition, the distance between an anaphor and its antecedent influences fixation times; when the antecedent is relatively far back in the text, fixations on the pronoun, as well as the next few fixations, tend to be longer (Ehrlich and Rayner, 1983; Garrod et al., 1994; O'Brien et al., 1997).

In addition to anaphoric referents, conjunctions as sentence connectors are also important devices to construct coherent text. In a written discourse, conjunctions signal the logical connections between ideas (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983; also see Geva, 1992) and also mark discourse structures and their functions, such as causal and temporal relations (Geva, 1983, 1992). Meyer (1977) pointed out that conjunctions help to make text organization explicit and coherent. As awareness of text organization is essential for text comprehension (Meyer, Brandt, & Bluth, 1981), conjunctions facilitate the instantiation of textual schemata (Kieras, 1985). The presence of conjunctions also help to direct reader's attention to important text information (Lorch & Lorch, 1986) and help reader to check information in memory (Spyridakis & Standal, 1987). This facilitated reading comprehension thus cost reader less reading time.

1.5.4 Reading Comprehension and Text Coherence

The comprehension of text, especially narrative texts, has been further investigated by the theorists who embraced construction-integration theory (Kintsch, 1988, 1998; also see Kintsch, 1974; Kintsch & van Dijk, 1978). They have argued that, during the comprehension of texts, readers construct a mental representation of the text as well as situations described in the text.

For example, van Dijk and Kintsch (1983) proposed that readers construct mental representations of (a) the text's surface structure, (b) the semantic meaning explicitly conveyed by the text or *textbase,* and (c) the situation described in the text, which is also called the *situation model.*

Within the frame work of situation model, researchers considered local and global text coherence as particularly important to text comprehension, i.e. to construction of mental textual representation (Kintsch, 1988) at surface form, the text base and the situation model level (Kintsch, 1994). Local textual coherence here refers to the fact that propositions of the textbase processed in working memory must share common arguments, while global coherence refers to the fact that the meaning of any textual information must match the situation model upon which the text's topic content bears.

Linguists have shown that causal connectives help construct a coherent text representation: the more causal relations/connectives readers identify in a text, the more coherent they perceive the text, and thus the easier they process the text and the better they comprehend and remember it (Van den Broek, 1988; Van den Broek, et al., 2001). They also suggest that that the connectives make the text more cohesive and structured by providing markers between sentences. In addition, connectives explicitly signal to readers that the sentences are connected with one another in a precise semantic manner. For example, causal connectives may incite readers to search knowledge in their long-term-memory in order to restore local or global text coherence. During this process, readers should be able to find the reason explaining the semantic connection between sentences, which facilitates their integration and comprehension of the text representation. This process by which related information is searched is referred as the mental generation of causal inferences.

Previous studies in narrative comprehension (Golding et al., 1995; Keenan et al., 1984;

Myers et al., 1987) examine the role of reader's search for causal relations in the construction of a coherent text representation and also explore the role of connectives in reading comprehension. Haberlandt (1982) found facilitative effects on reading time with causal conjunctions *therefore, so, consequently* in connective-present sentences versus no-connective sentences. The findings indicate that target sentences preceded by a connective result in faster reading times than unconnected sentences. Trabasso et al. (1984) distinguished between short term and long term connectivity underlying the construction of coherent relations. The former one, derived from linguistic cohesive devices, generates local coherence, while the long-term connectivity is constructed when readers draw on their world knowledge to construct the causal connections that represent the information of narrative texts. Therefore, readers construct a coherent text representation that is primarily driven by an intuitive expectation of satisfying cause-effect relations. Keenan et al. (1984) also explored the impact of causal relations on text comprehension, suggesting that causal connectivity between sentences plays an important role in the construction of coherence relations. They claimed that a coherent text interpretation emerges from knowledge-based relations constructed during the process of inter-clause integration. Results of their study partially confirm that inter-clause integration entails the construction of knowledge-based relations such as cause-effect sequences (Keenan et al., 1984).

In addition to causation, researchers suggest that situation models, at least in narrative texts, consist of another four dimensions including time, space, motivation and protagonist. These dimensions also help to construct text coherence. Zwaan, Magliano, and Graesser (1995) reported that coherence breaks on situational dimensions affect reading time. They found that the temporal and causal inconsistency in a text lead to significant increases in readers' sentence reading time for short stories. This finding indicates that the break of text coherence makes it

difficult for readers to integrate upcoming information into the evolving mental representation. The study of Zwaan, et al. (1998) expanded the finding of Zwaan, Magliano, et al. (1995) by exploring all five dimensions of the situation model. This study found that people monitor the coherence continuity on multiple situation dimensions. As a result, reading time was increased by the discontinuity of any/all of these five situational dimensions.

The findings from the literature imply that reading time should be a robust indicator of text comprehension. Within the context of essay grading, it is thus plausible to predict that raters' sentence reading time for an essay would increase if the text has a high density of the following features: 1) words with long word length (or more syllables) and low-frequency; 2) sentences of syntactic anomaly such as the subject-verb disagreement; 3) sentences containing multiple clauses and s complex sentence structure (long sentence length); 4) inconsistent anaphoric referent; 5) insufficient use of sentence connectors; 6) inconsistency in the text representation of time, space, causation, motivation and protagonist.

CHAPTER 2

PROPOSAL

Writing test as a performance-based language assessment is a multifaceted entity involving the interaction of various factors, among which essay rater's subjective judgment has a great impact on essay score, thus influencing the validity and reliability of a writing test. Rater's scoring procedure, however, is not an objective and error-free process. It is the final output of a series of scoring behaviors including reading, text comprehension, evaluation, and scoring decision making. Rater's reading comprehension, as an inseparable component of the rating process, is in fact the prerequisite of a reliable score judgment. In other words, a writing test is not able to reliably measure the targeted writing ability unless raters fully comprehend the writing responses.

Despite rater's impact on test validity and reliability, traditional methods for the study of writing tests are based solely on test score, which is normally an interval or ordinal measurement of test-takers' ability as defined by the test construct. The current study expends the scope of rating study into raters' scoring behaviors and their reading comprehension. In the proposed framework, rater reliability thus can be redefined as the desired set of scoring behaviors; and test validity should also be assured through a set of scoring behaviors authentic to what test-makers would expect from raters.

In the current model, the structure of scoring behaviors in a writing assessment can be simplified into three levels, as seen in Figure 2.1. On the top of the scoring pyramid is the final output of the rating process - the score of a test, which is readily observable for most types of writing assessment. The traditional methods focus only on the score level information by

correlating it with various rater attributes, text attributes and test-taker attributes. Beneath the final output of test scores lies the scoring behavior, which largely governs the quality of scores. Since raters in a writing assessment are also text readers, their reading comprehension as a scoring level is as fundamental as their scoring behaviors. Many researchers have realized the importance of integrating these lower level scoring procedures into the models for writing assessment, however, the limitations in previous methodologies have not been completely overcome.

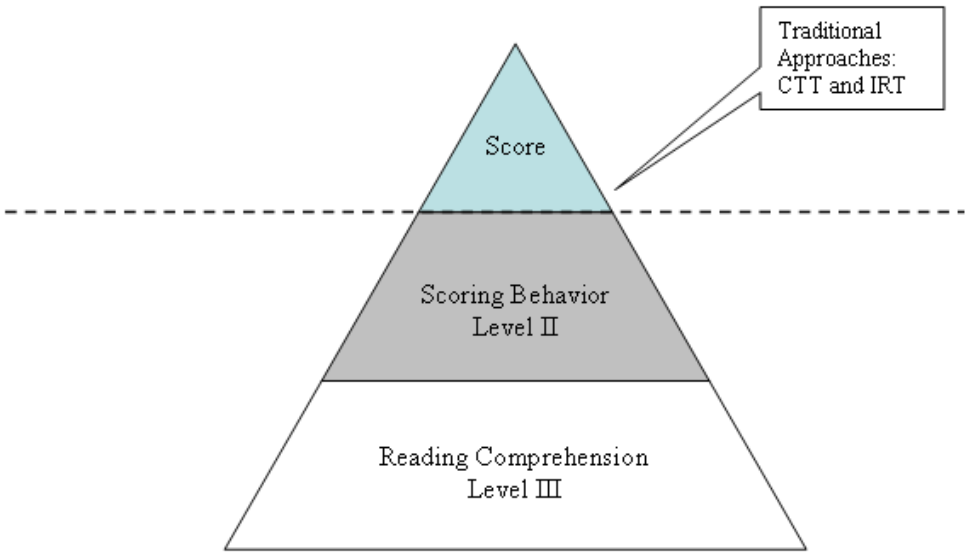


Figure 2.1. The Structure of Raters' Scoring Process in Writing Test.

Rater's scoring behavior (Level II) includes a spectrum of activities, most of which are not easily observable. This is why previous researchers had to limit themselves to the final score output. By designing a new data collection instrument, the present study is able to record and analyze raters' reading pattern, evaluation process and their decision making process. With the renewed framework of analysis, the current investigation expands the definition of rater reliability to the degree to which rater's actual scoring behavior coincides to the scoring behavior

defined by rating rubrics. This definition is different from the more traditional and statistical interpretation of rater reliability, but the researcher argues that it is more consistent with the understanding of rater reliability by test practitioners, policy makers and researchers in psychology and applied linguistics. In addition, the current framework reinstates that the validity of writing assessment depends on the reliability of essay rater. If raters are not reliable, even if the test itself is appropriately designed, the results of the tests may be invalid. For example, if the test is designed to evaluate one type of writing skill, while the rater evaluates the test based on irrelevant skill sets, the validity of this test is seriously eroded. Again, the connection between rater's scoring behavior and test validity is realized through the prescribed rating rubrics.

This study also points out that a seemingly accurate score assignment itself does not insure validity and reliability, even if an independent argument of its correctness is available. If rater's reading comprehension is flawed, even if the scoring behavior is a correct reflection of rating rubrics, the final score assignment might be biased as well. Since reading comprehension is a psychological process which cannot be directly observed, the researcher investigated this process through inferences made from raters' reading patterns. Although the current study proposes a measurement model, it is actually based on the literature of reading comprehension (Level III). In the current study, the researcher intends to integrate previous findings on the study of reading comprehension into the current test model, as well as design new methods to further explore the reading comprehension patterns of essay rater.

Based on the previous arguments, the current study proposes a behavioral model for writing performance assessment. This model defines and explores rater reliability and test validity via the interaction between text (essays written by test-takers) and rater. Instead of indirectly approaching the success of such an interaction through essay scores, the new testing

model directly measures and examines the success of raters' behaviors with regard to essay reading and decision making. Because it reveals the interactional nature of a performance test, this new model is named as the Interactional Testing Model (ITM). The general framework of ITM can be generalized into a broader test context as displayed in Figure 2.2.

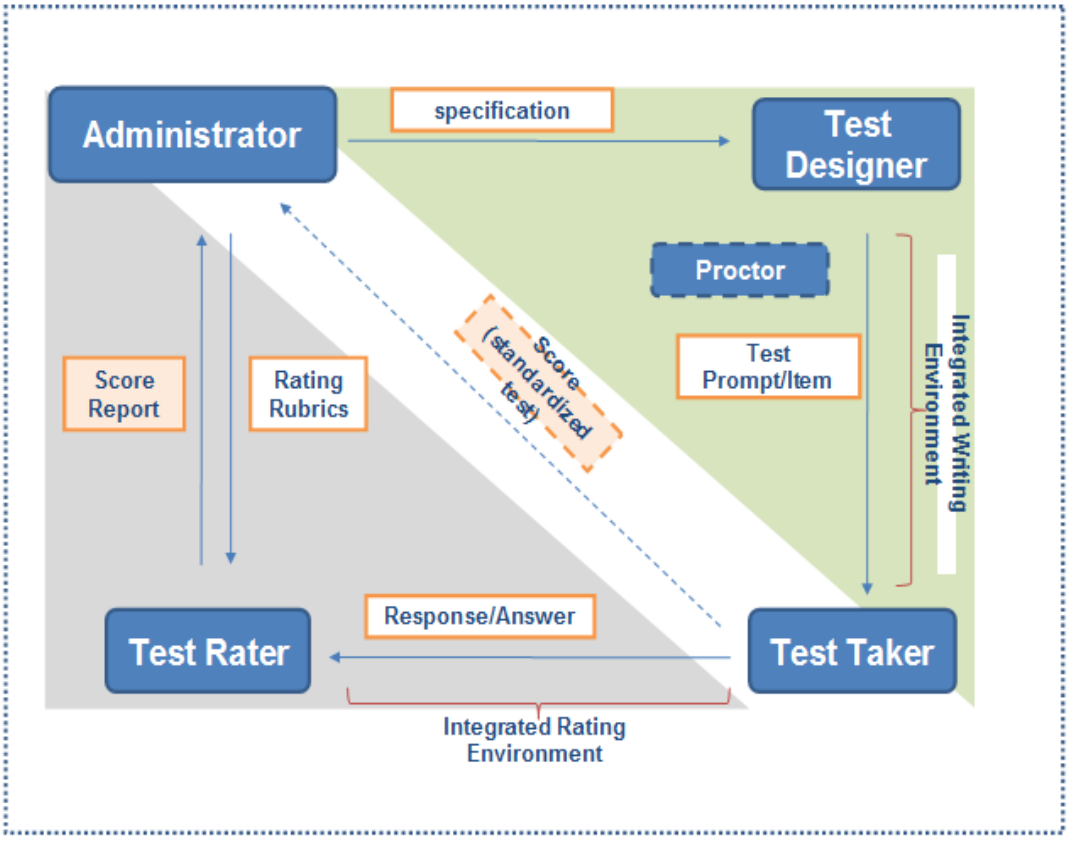


Figure 2.2: The Structure of the Interactional Testing Model

The framework of ITM considers the whole process of testing as the interaction between various test stakeholders. The interaction between test maker and test taker is realized directly through test and indirectly through scores, with essay rater as the media; on the other hand, the interaction between test taker and essay rater is realized through essays. In this study, the issue of test validity is revisited indirectly through the investigation of rater reliability. Raters' scoring

processes are examined through three aspects including scores, scoring behavior and reading comprehension.

This new testing model, however, does not attempt to reject the traditional statistical methods such as IRT and CTT. Instead, the current proposal is that ITM framework is a supplement of IRT and CTT since it expands into a realm of new phenomenon that is beyond the current consideration of traditional methods.

2.1 Research Hypotheses

In order to examine rater's decision making process in the EPT writing test, four research hypotheses are proposed in this study.

Hypothesis 1: *A high reading digression rate and a low reading rate indicate an engaged reading comprehension process during essay grading, hence these indices are positively associated with rater reliability in a writing test.*

Hypothesis 2: *If there is an interaction between rater and essay writer, raters' scoring decision is associated with essay features.*

Hypothesis 3*: Rater decision making is reflected not only in their score assignment, but also in their scoring behaviours such as sentence selection, verbatim annotation and comment.*

Hypothesis 4: *Raters not only have an agreement on score assignment, but also share a common scoring focus when evaluating writing qualities.*

CHAPTER 3

EXPERIMENTAL DESIGN

The proposed ITM framework in this study is adopted to investigate rater's decision making process when grading ESL essays. This study looks into the impact of rating on the construct validity of the EPT test at UIUC. The purpose of the current study is thus to evaluate if the Semi-Enhanced EPT measures the target construct and if raters' scoring behaviors are consistent when they read and grade the texts.

3.1 Research Context

The EPT at UIUC is a year-round test given to all incoming international students whose TOEFL or IELTS scores are at or below the campus or departmental cutoff scores: 610 for paper-and-pencil TOEFL, 253 for computer-based TOEFL, 102 for internet-based TOEFL, and 6.5 for IELTS. As the primary tool of post-matriculation screening, this test is used to place international students into appropriate ESL writing and/or oral courses.

The EPT consists of two parts: a writing test and an oral interview. The purpose of the oral interview is to identify students who need to take an ESL pronunciation course to succeed in their study at UIUC and then place them into the appropriate ESL pronunciation courses. In the oral test, students are interviewed individually by an experienced ESL teacher. As the present study focuses on the EPT writing test only, the oral interview subtest of EPT will not be discussed. In this paper, the EPT test only refers to the writing subtest of EPT.

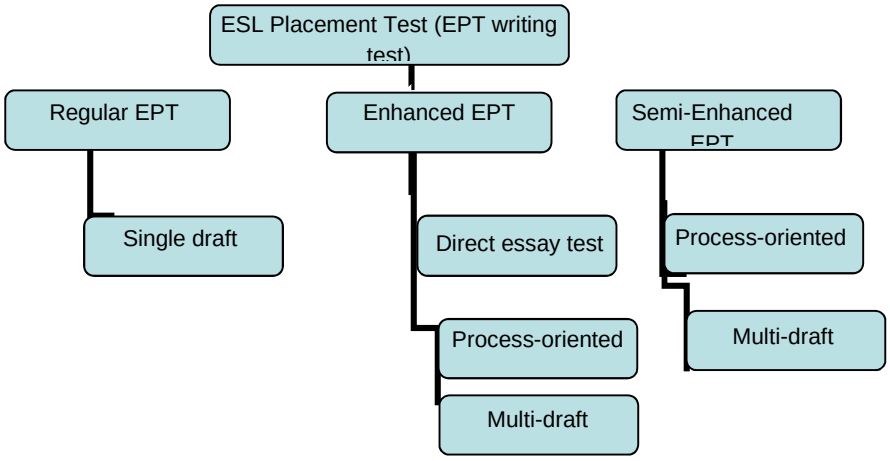


Figure 3.1: Three versions of EPT writing tests.

The EPT writing test is an integrated, English for Academic Purpose (EAP) placement test (Pyo, 2001, also cited in Lee, 2005). There have been three versions of the EPT writing test, including the "Regular EPT", "Enhanced EPT", and "Semi-Enhanced EPT (SEEPT)" (Figure 3.1). The regular EPT is a 50-minute single-draft writing assessment. Students are required to watch a videotaped lecture, read an article related to the content of the video lecture, and then write an essay to demonstrate their understanding of the stimuli materials. The Enhanced EPT is a day-long process-oriented multi-draft essay assessment. it is a workshop-based essay test that consists of a morning session and an afternoon session: in the morning session, the proctor introduces the writing topic and facilitates a brainstorming and group discussions among examinees, who afterward watch a video lecture, read a related article, and write their first draft; in the afternoon session, test takers produce the finalized essay based on their self-evaluation and peer feedbacks on their first draft. By having examinees fully engage in the writing process, this test is expected to elicit a comprehensive range of writing abilities and to obtain writing

performance samples that are a more accurate reflection of examinees' writing instruction needs (Lee, 2005).

The current version of the EPT, SEEPT is also a process-oriented multi-drafting writing assessment approximately four hours in length. This integrated writing test requires students to produce an academic essay based on the information received from a reading passage and a short lecture. After the mini lecture, the test proctor will provide a scoring rubrics which inform student the required features that their essay needs to contain: 1) a clear organization of introduction, body and conclusion for an argumentative essay; 2) explicitly connected ideas; 3) ideas supported with information from BOTH the lecture and the article; 4) accurate understanding of BOTH the lecture and the article; 5) identified source of information; and 6) grammatical accuracy.

In the SEEPT, the video tape lecture is replaced by a class lecture delivered by a teacher/proctor, who is an experienced ESL instructor at UIUC. After the lecture, this teacher will lead a class/group discussion to help test takers to comprehend the writing topic and the stimuli materials. The purpose of this change is to mimic the lecture-discussion interaction between professor and students in the real world classroom, thus providing a more realistic context for the assessment of EAP. In the SEEPT, examinees first read an article on a given topic and then attend a lecture and discussion as a whole class. After the discussion session and the explanation of scoring rubrics, students are required to produce an outlined first draft of their essay based on a writing guideline provided by the proctor. The purpose of this outlined draft is to help students to organize their thoughts and formulate the overall structure of their essay. After the first draft, test takers pair up and peer evaluate their partner’s writing. Based on the outline and the feedback from their peers, examinees take another hour to produce the final draft of their

writing response to the essay question.

The SEEPT has the following advantages that elicit the best possible performance from test takers. First of all, it constructs a realistic context to assess examinees' EAP. It also ensures that examinees understand the essay topic and enables them to employ support materials during the test. Compared with EPT and EEPT, the SEEPT also employs the facilitative activities and focuses on examinees' writing process, while it requires less technical support and takes less time. As the video lecture has been replaced by a classroom lecture, the SEEPT can be administered in most classrooms on campus. Test takers also find this test version more time-efficient. The EPT registration of fall 2006 indicates that most test takers preferred SEEPT to EEPT when the pilot SEEPT test was advertised on the registration website. The SEEPT has replaced the EPT and EEPT to be the only available test format since the summer of 2007.

The writing responses in three versions of EPT are graded based on the same rating rubrics that measure the same constructs. This rating rubric adopts the concepts and features of holistic scoring; however, it does not encourage raters to assign a score based on their general impression of a writing sample. Instead, raters are required to evaluate writing at different performance levels in explicit scoring criteria. In the current EPT rubrics, writing proficiency is measured by a four-point scale in four rating dimensions, including *Organization*, *Development*, *Grammar and Lexical Choice*, and *Plagiarism*. The development of scoring rubrics is consistent with the multidimensional nature of language proficiency (Bachman 1990).

Each of these four dimensions is divided into four levels with score points ranging from 1 to 4. The writing responses are graded by experienced teaching assistants (TAs) in the Division of English as an International Language (DEIL) at UIUC. In the operational EPT scoring, all raters are instructors of ESL courses and have attended mandatory writing rater training led by

the ESL TA supervisor. Each essay is read by two raters and the final score is the one two raters agree on. In case of extreme score differences (more than 1 score point), the essay is given to a third reader, and the two scores which are closest to each other are used to determine the final score.

3.2 Research Methods

3.2.1 Participants

Twelve EPT raters participated in the present study (nine female and three male). Ten of them are international graduate students in the MATESL program and two are native speakers. All participants are fluent in English reading and writing, therefore, their language proficiency should not affect their reading comprehension of EPT essays. All these participants had taught ESL writing service courses at UIUC, but only seven had prior experience of operational EPT essay grading. Those experienced EPT raters had attended the operational rater training session and rated EPT essays using the current rating rubrics. The new raters, including the two native speaking graduate teaching assistants, had never graded EPT essays before the data collection; yet they were quite familiar with the rating scale, the essay prompt, and the level of students' writing among test takers as they were teaching the same population in their ESL classes. Since the current study does not emphasis in the language aspect of the EPT test, rater performance would not be affected by their language background. These twelve raters also shared similar professional backgrounds. On average, raters had learned English for over 10 years and had been teaching English for over 3 semesters at UIUC. Before they were admitted by the MATESL program, all raters had taught in an ESL/EFL context for at least more than one year.

The major reason to choose these twelve participants is that this group represents the major background of typical EPT raters and thus constitutes a sample that is representative of the population to which the study is intended to generalize. Though there are only two English native speaking raters included in this study, the rater group was viewed as representative due to the limited number of native speaking raters at DEIL.

This study adheres to all rules set forth by UIUC and College of Education for the use of human subjects in research. The original research plan was submitted to the UIUC Campus Intuitional Review Board (IRB) and received approval before the data collection. The researcher made sure that the confidentiality of all participants throughout the course of the study and thereafter. All participants were informed in the study consent form (see Appendix C) that their answers will be kept confidential. All participants were fully informed of the purpose of the study, the potential benefits of the study, the anticipated use of the data, and their rights and responsibilities as study participants. They were informed that they have the right to refuse to participate in the study or to end their participation in the study at any time. All participating teachers were given an ID number and no identifying information was included in the database that contains their grading responses. No individual responses were attributed to an individual participant by name or by any other way that they can be specifically identified. This database was password protected and accessible only to the researcher. This database was not being stored on any network space.

3.2.2 Materials and Procedure

a. EPT essays

20 SEEPT essays were randomly selected as secondary writing data from 2007 EPT

administrations. These essays had previously been sanitized by removing examinees' background information including their name, major of study, university ID number and their student status. Each essay was referred in this study by a file name consisting of its test date and a serial number. These experiment essays are stratified samples that represent all four levels of proficiency among EPT examinees, ranging from grade 1—too low, grade 2—ESL 500, grade 3—ESL 501 and grade 4—exempted.

b. Rating rubrics

The previous EPT rating rubrics was developed by Lee (2002) as a holistic scoring scheme with four categories: 1) *Organization* evaluates if a writing response has a clear structural organization including introduction, body and conclusion; 2) *Development* examines the development of writer's thesis statement; 3) *Grammar and lexical choice* looks into the linguistic feature in the writing responses; and 4) *Plagiarism* dimension tells if test takers appropriately document the source materials as the supporting evidences. For each category, there are four full letter scale levels from 1 to 4 (see Appendix B).

c. Rater Training

All raters participated in a 60 minute training session at fall, 2007, which took place in a computer room in the Foreign Languages Building of UIUC. The training session was delivered to all raters by the researcher, using the same training materials for demonstration and practice. At the beginning of the training session, raters were given a copy of the SEEPT reading passage of the target topic, related lecture notes and the SEEPT rating benchmarks. Raters then had 15 minutes to get familiar with the topic of the selected SEEPT essays. After that, the researcher led a 10 minute review session to go over the rating rubrics and clarify the rating scales. A brief description was also given on the definition of the four scoring criteria. After the review of rating

rubrics, a demonstration tour of the rating instrument was given to each rater to teach them how to use a computer-based scoring interface to grade SEEPT essays and what the grading requirements were. First of all, each rater read a handout of the interface user manual. When they finished reading, the function of each section on the interface was explained and demonstrated by the researcher. Raters then made their own practice on a computer by grading four stratified sample essays (the same across raters). After a five-minute break, raters discussed with the researcher the ratings they had just given and had a short Q and A session about the function of the interface. After the discussion, each rater opened a new interface on their computer and started grading the experiment essays preloaded in the scoring engine.

3.3 Instrument: The Integrated Rating Environment

The major difference between previous studies on rater effect and the current research is that a computer based rating interface is designed for this study to deliver students' writing samples and collect raters' scoring data during their decision making process. This rating interface is a Geographical User Interface (GUI) written in Python with the Tkinter package. It can be run on any Windows operating system. The purpose of the rating interface is to automatically detect raters' scoring event and process all grading records including score assignment, reading speed, reading regression, scoring comments and sentence annotation made by each rater. This rating instrument addresses the rater-text interaction in this study and also allows raters to read, grade and answer post-rating questionnaire on the same computer interface, therefore, the current rater interface was named the Integrated Rating Environment (IRE). Compared to eye-tracking devices and retrospective data collection using paper surveys, the IRE is a more cost effective tool that is able to capture raters' reading activities and automatically

generate data for analysis.

The scoring page of the IRE can be divided into six major sections, including file buttons, a search engine, a scoring section, radio buttons, a timer and a text window (figure 3.2).

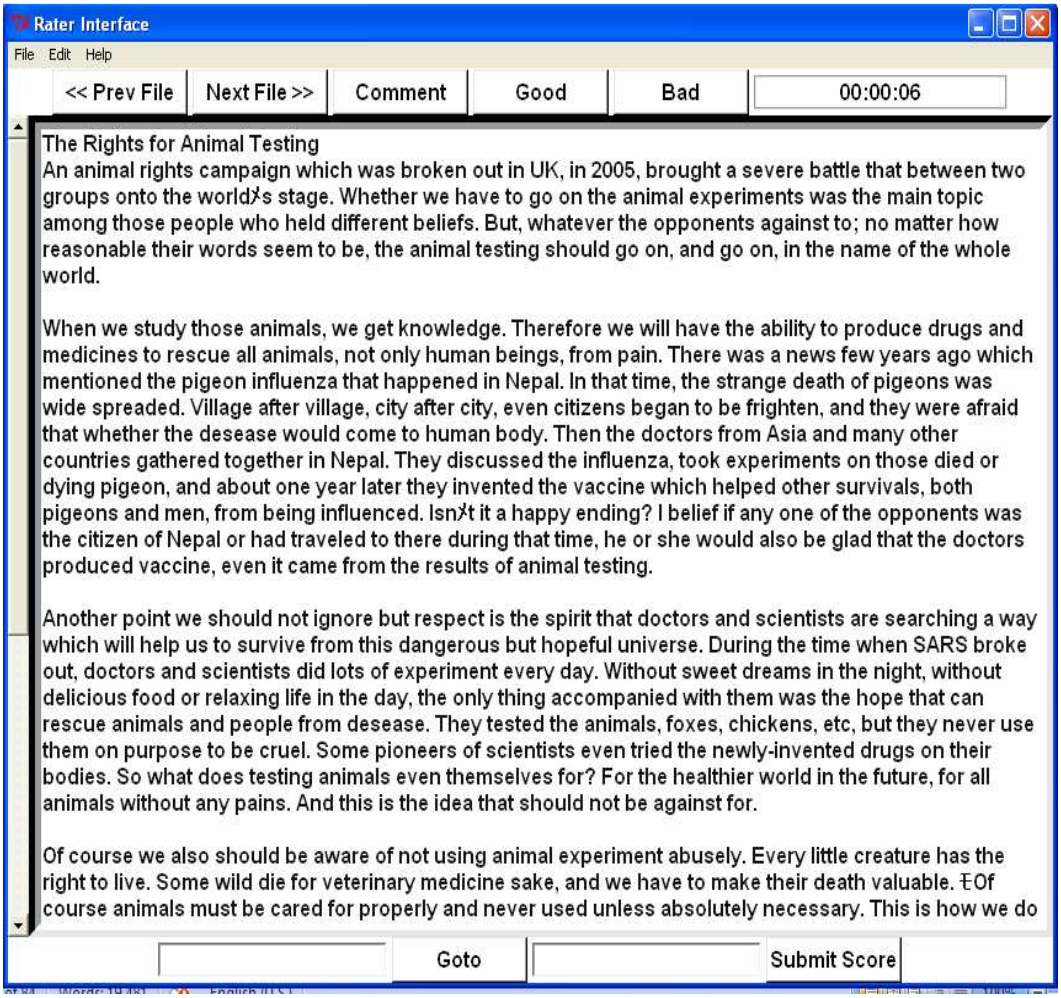


Figure 3.2: Front Page of the Integrated Rating Environment.

These six sections are associated with particular functions: 1) the text window is used to display the written samples from examinees. In order to avoid the halo effect in rating, only one essay appears on the window at one time. When the IRT is open, the color of the sample script gradually fades away in 30 seconds so that the script will be too light to read. In order to read on the text window, raters were required to use the mouse to highlight the sample script as they read

the essay. They were also asked to annotate sentences or phrases from sample writings as either positive or negative scoring evidences that help them to assign a score. When leaving verbatim annotations, raters highlighted sample scripts and clicked either the “Good” or “Bad” citation button to mark them as positive or negative scoring evidence. After doing so, the annotated sentences would be marked in the essay on the text window (Figure 3.3). On the instruction page, the text window also gives raters a brief description about how to use the interface.

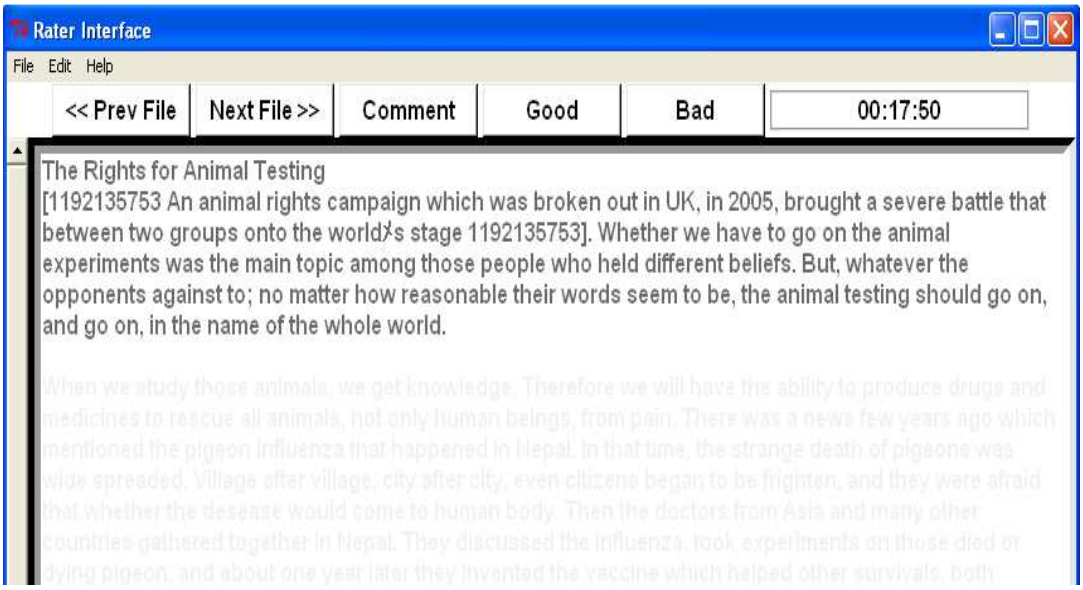


Figure 3.3: The display of annotated sentences in the text window.

There are five radio buttons and a clock above the text window: the clock records the total grading time for each rater. Raters used the “Prev File” and “Next File” buttons to go back to the previous essay or move to the next essay. The “Good” and “Bad” annotation buttons were used to assign sentences/phrases as raters’ scoring evidence. If raters would like to leave any comments or feedbacks during grading, they clicked the “Comment” button, typed their comments in the comment window and inserted the comments into the original text by clicking in the text and then pressed the “Insert Comment” button. A sample lay-out of the comment

window can be seen in Figure 3.4.

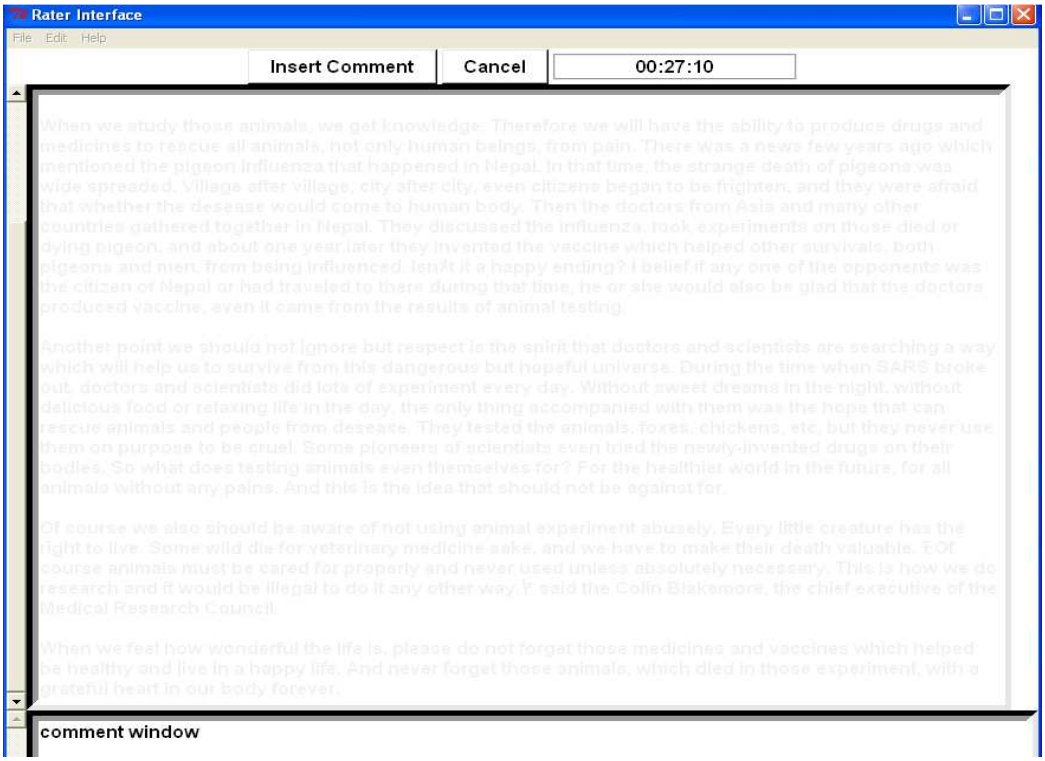


Figure 3.4: The layout of interface before rater inserts a rating comment.

Below the text window of the IRE, there is a search engine and a scoring section. The former helped raters to locate a particular essay by searching the serial number assigned to that essay. The scoring section was used to assign a grade to the present essay. The scoring scales ranges from 1 to 4, which stands for different performance levels in EPT writing section. Only one letter grade was allowed in this study.

On the top of the interface, there are another 3 file buttons include: “File”, “Edit” and “Help”. The “File” button provided options for raters to hide their comments or scoring annotation in the original text, or helped raters to check the comment and citations without reading through the whole passage. (Figure 3.5)

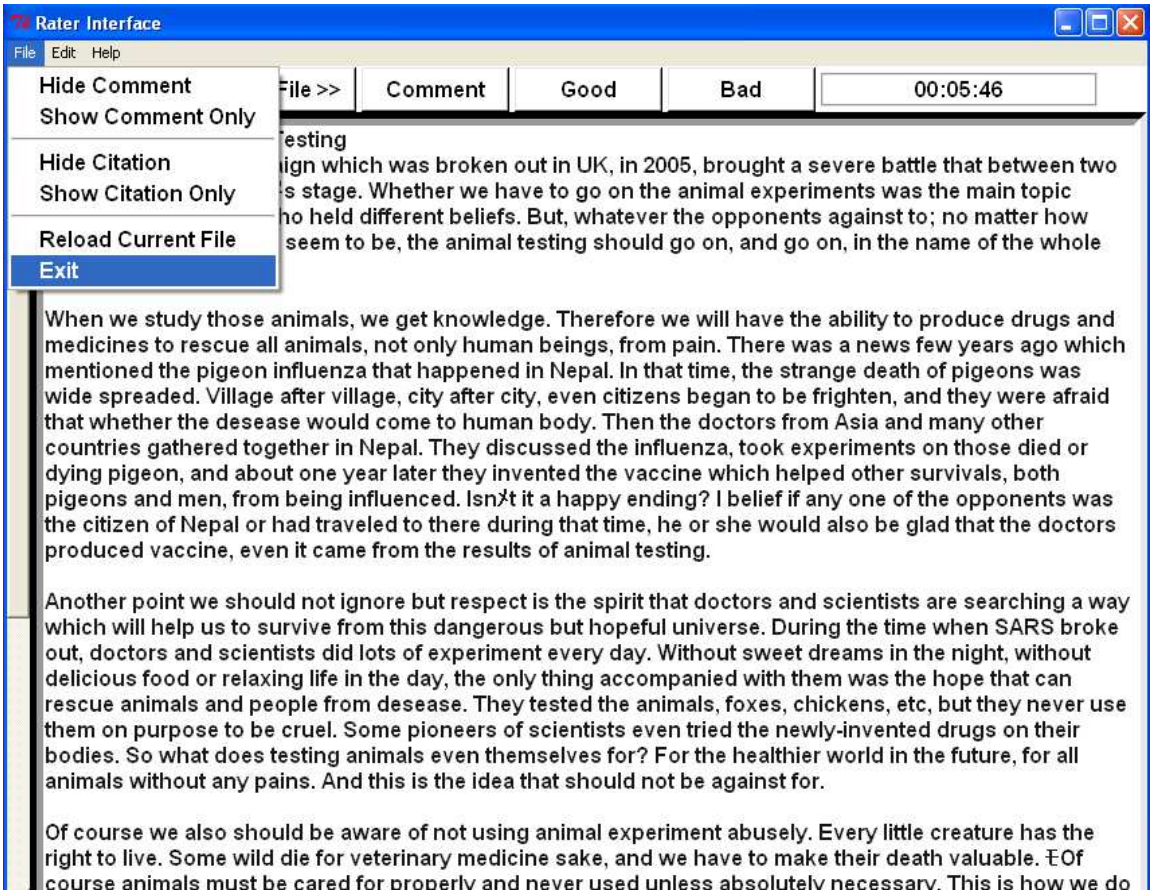


Figure 3.5: The function of the "file" button.

The "Edit" button can be used to delete annotations or comments that raters made. If raters left an inappropriate comment by mistake, they could use this "Edit" button to remove the record that they just made. This button also provides the option that raters may remove all the scoring annotations or comments for a particular essay and re-do the scoring. (Figure 3.6)

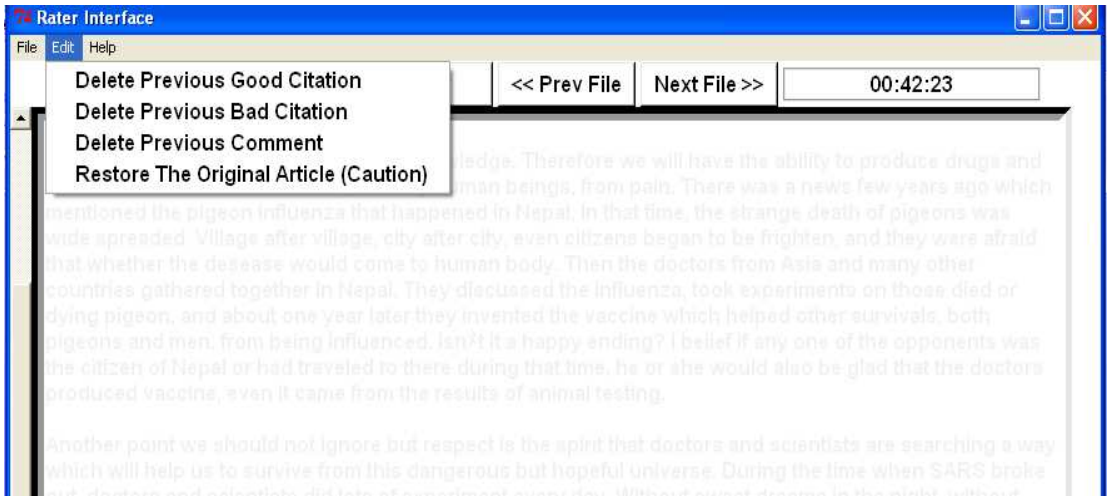


Figure 3.6: The function of "Edit" button.

When raters finished grading, they were directed to four scoring questions by clicking the "Next File" button. (Figure 3.7)

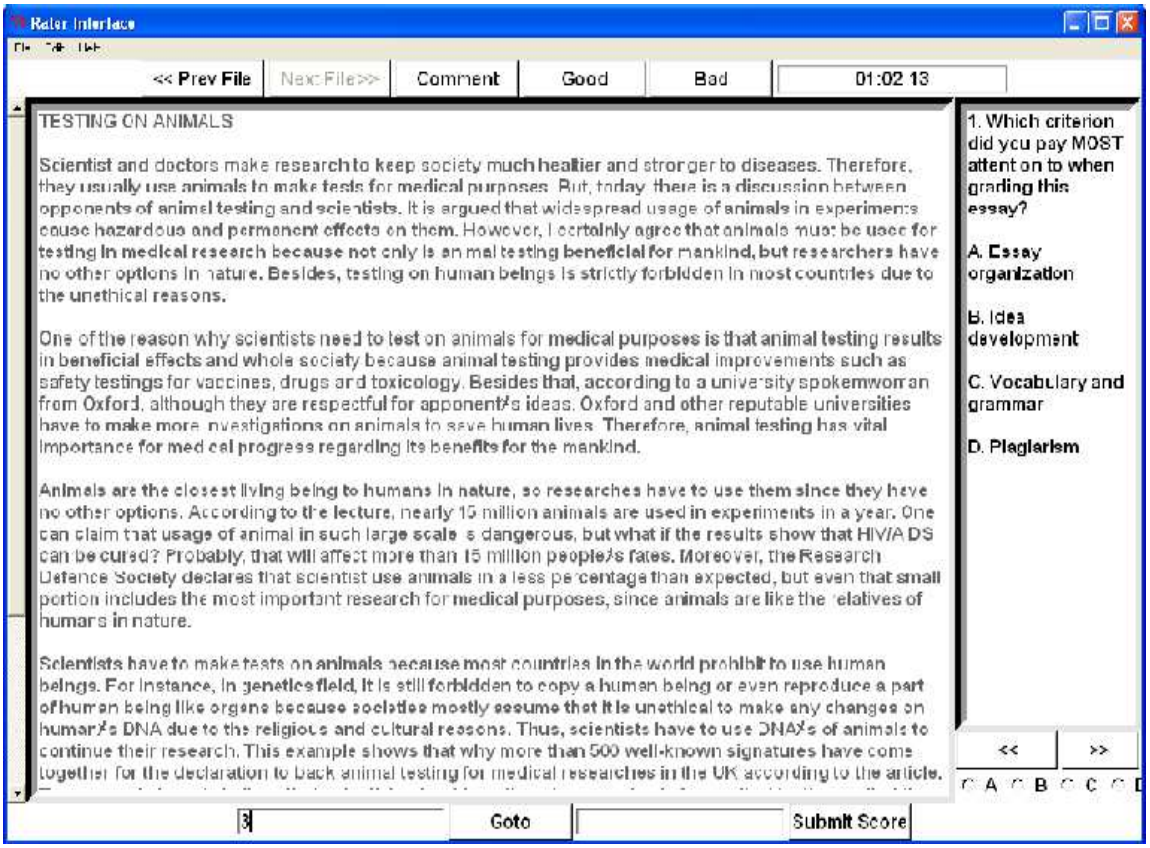


Figure 3.7: The display of the essay question window.

There are two multiple-choice questions and two short-answer questions for each essay. These four questions are the same across all essays. To answer the multiple choice questions, raters chose one of the four radio buttons A to D in the question window and click the arrow button to move to the previous or next question. To answer the short answer questions, raters first clicked the "Answer" button in the question window, typed their answers in the pop-out answer window and then clicked "Submit" button to turn in their answers. When they were done with all four questions, raters moved to next file by clicking the right hand arrow button.

All of these scoring events were automatically recorded by IRE and a timed scoring log was generated for each rater. This log displays raters' reading behaviors by specifying when a rater started and finished grading and also what particular script this rater was reading at a

particular time. In this case, the pattern of raters' text reading can be estimated based on their reading speed and reading regression. In addition to raters' reading pattern, this log also provides temporal and spatial information of raters' scoring comments and sentence citations. (Figure 3.8)

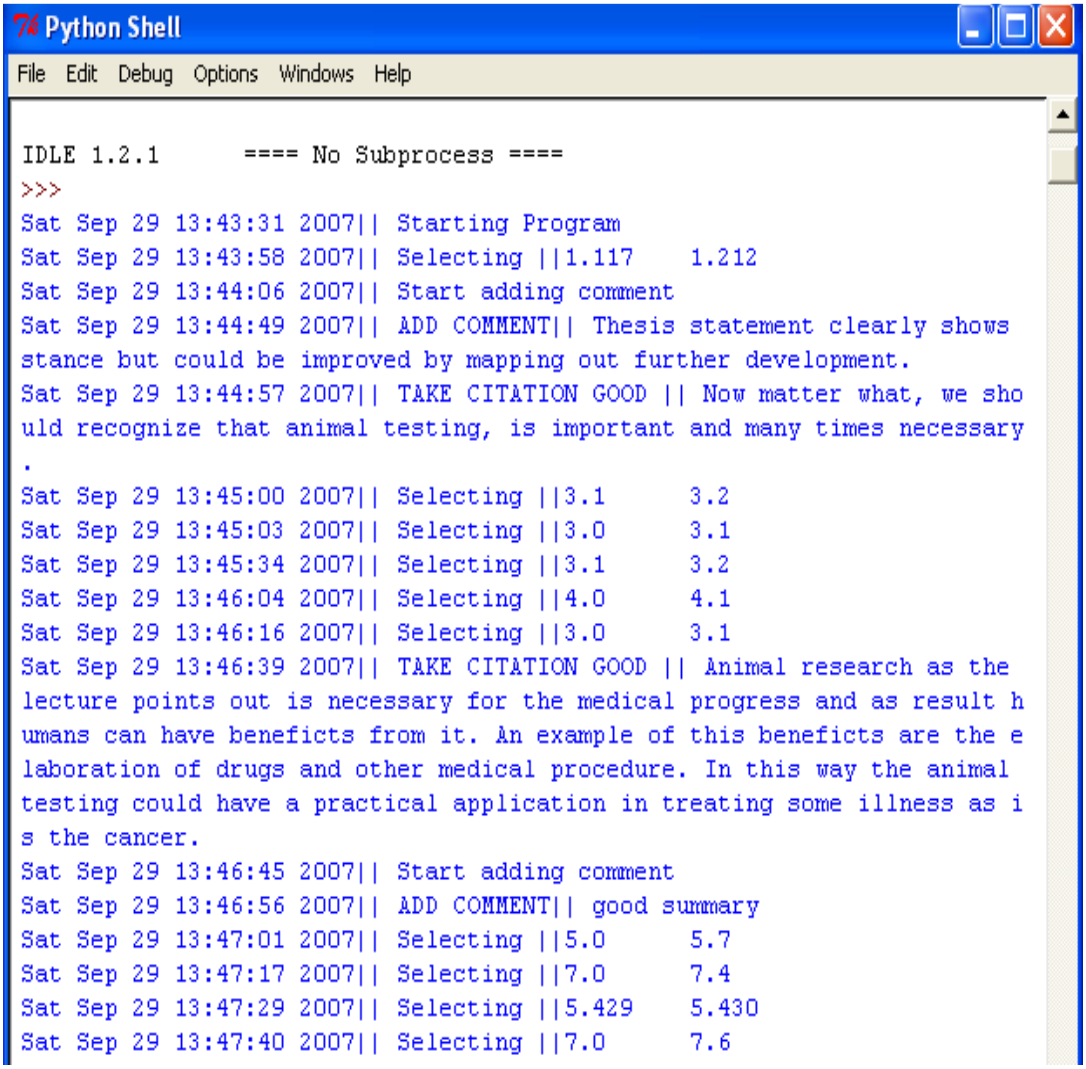


Figure 3.8: The layout of a typical rater's' scoring-event log.

Each individual rater was given a grading folder which consists of a copy of the rating interface with assigned writing samples preloaded into the data engine and a text file of user's guidebook of the IRE. A 15 minute demonstration session was also given to all raters on how to use the interface on their own computer. To start grading, raters were required to copy the rating

interface onto the desktop of their own computer. Raters' reading behavior or their scoring record were automatically detected and saved in the engine whenever raters made a scoring event, such as highlighting a sentence or clicking a radio button. If they would like to stop in the middle of their grading, raters was informed to close interface by clicking "Exit" in the "File" button. When the rating was completed, raters were asked to compress their scoring folder onto the desktop of their own computer and uploaded the zipped file to a shared website.

3.4 Procedure

The current research took in a computer room at UIUC. The rater training and essay grading were held in the same room. The researcher first sent an email invitation to all of the current EPT raters to explain the content of this study and ask for their participation. The first twelve raters who contacted the researcher to confirm their participation were selected. Before the experiemental session, raters participated in a 60-minute training where the participants were taught how to use a rating interface to grade EPT essays and make practice on a group of sample essays.

After the training session, each participant graded 20 EPT essays which were identical across participants. During their grading process, raters were required to annotate sentences/phrases from the sample essay as the evidence of their score assignment. They were also asked to leave comments and answer rating questions on the IRE. Raters' scoring record and decision making process were monitored and further analyzed by the rating instrument in the present study.

3.5 Data

The data collected in the current study consist of three parts. The first part is the sanitized writing responses from previous EPT test takers. The researcher uploaded writing samples into a computer-based rating interface and assigned each examinee a unique ID number that appeared on the rating interface.

The second part of the data is raters' scoring records collected by the rating environment. These scoring records include raters' scoring scale choice, time of each rating event, their scoring annotations, comments and their responses to survey questions. The ID number associated with each examinee/rater were used as the file name to differentiate the source of rating records. During the study, only this number was referred to instead of any personal information of the participants.

The third part of the data was collected from participants' survey questionnaires after their grading session. As rater reliability may also be affected by raters' professional background, a survey questionnaire was designed to collect raters' background information with regard to their ESL/EFL teaching experience, instructional focus and also their essay scoring experience. This questionnaire was also used to elicit from raters their reflective feedbacks on the training session and their rating process. This three page questionnaire consists of 9 Matrix Questions and an open end question (see Appendix A). The questionnaire was designed in this way to be more user-friendly to the respondent and also to assure the comparability and comprehensibility of responses by eliciting both objective and subjective responses. This questionnaire was emailed to each rater after the experiment session. They were asked to upload their anonymous questionnaire onto a shared website to assure that all survey questions were honestly answered.

3.6 Measurements

In the current study, the researcher measured several important entities representing essay features and raters' dynamic scoring process. Three groups of variables were measured in this study. These variables include essay features, raters' reading comprehension and their essay scoring behaviors. The interaction between these three categories were analyzed to test related research hypotheses in this study, thus helping us to understand raters' reading comprehension and decision making process when grading ESL essays.

Reading Pattern: raters' scoring events and their reading responses were automatically monitored and recorded by the IRE.

1) Readers' total reading time and letter-per-second reading rate for each essay were recorded by the interface.
2) Raters' go-back rate within and across paragraphs.
3) The time-by-location information of raters' sentence selection in each experiment
4) essay, including when, where and how many times raters regress to a previous sentences during reading. This information was monitored automatically via the mouse click during sentence selection.

Reading Comprehension and Scoring Behavior.

1) The time-by-location information of raters' verbatim annotation as both positive and negative evidences of their scoring decision. The temporal and spatial information of raters' annotation was recorded when raters highlighted the selected sentence and click related category button (Good or Bad).
2) The time-by-location information of the raters' comments. The interface recorded when and where raters inserted comments and how much time it took them to formulate their

comments.

3) The letter grade score for each essay.

4) Raters' responses to four scoring questions after grading each essay. Their answers to two multiple choices questions and two short answer questions were extracted from the interface, as well as their response time.

5) Raters' answers to a survey questionnaire after the experiment session. Information with regard to raters' self-reported teaching and scoring experience were collected from the questionnaire.

Essay Features: The experiment essays were processed and analyzed by Python and SAS.

1) Word frequency. The experiment essays were processed by Python to examine their average word frequency.

2) Essay length. The total characters in an essay were calculated by Python as the indicator of essay length.

3) Total number of subject-verb mismatch at sentence level for each essay was estimated as the indicator of syntactic anomaly.

4) Total number of clauses in each essay and letter-per-sentence sentence length was calculated by Python as the indicators of syntactic complexity.

5) The total number and location of inconsistent anaphoric referent and the total number of tense shift in each essay were calculated as indicators of discourse incoherence.

6) The density and word frequency of sentences connectors in each essay were calculated by Python as indicators of discourse coherence.

CHAPTER 4

RESULT

4.1 Rater's Reading Pattern

Results from the current study indicate that essay raters had different reading speeds during text reading. Some raters' reading rates substantially deviate from the group mean. Grader's letter-per-second (LPS) reading rate for each essay is demonstrated in Table 4.1, which displays that the mean reading rate varies across rater. Compared to the group mean reading rate 17.58 lps, the mean reading rates for some raters, e.g. rater 1, 4 and 7, are remarkably higher. Rater 5 and rater 9, on the other hand, had surprisingly low reading rates at 8.74 lps and 8.73 lps, respectively.

Table 4.1: Raters' letter-per-second reading rate.

ID	N	Mean	StdDev	Min	Max
1	19	25.28	7.63	15.50	42.37
2	19	11.76	4.02	5.84	17.54
3	20	20.94	4.80	11.79	30.38
4	18	26.43	14.23	10.10	49.29
5	20	8.74	4.32	4.96	20.62
6	18	18.00	5.49	9.81	27.65
7	20	24.93	9.94	9.04	44.09
8	20	21.16	5.07	11.71	30.33
9	20	8.73	4.57	2.49	22.26
10	20	12.37	3.91	4.06	20.03
11	20	16.81	4.79	9.64	26.00
12	19	15.80	8.42	7.63	31.85

Note: * N is not always 20 as some raters accidentally skipped essays.

In order to get a better understanding of the normality of rater's reading speed, the LPS reading rate is transformed into word-per-minute rate (WPM). Using data from the UDHR in

Unicode database[1], English has an average word length of 5.10 characters. The estimated WPM reading rates for twelve participants are displayed in Table 4.2.

Table 4.2: Rater's word-per-minute reading rate.

ID	N	Mean	StdDev	Min	Max
1	19	300.94	7.63	182.33	498.47
2	19	138.35	4.02	68.71	206.35
3	20	246.35	4.8	138.71	357.41
4	18	310.94	14.23	118.82	579.88
5	20	102.82	4.32	58.35	242.59
6	18	211.76	5.49	115.41	325.29
7	20	293.29	9.94	106.35	518.71
8	20	248.94	5.07	137.76	356.82
9	20	102.71	4.57	29.29	261.88
10	20	145.53	3.91	47.76	235.65
11	20	197.76	4.79	113.41	305.88
12	19	185.88	8.42	89.76	374.71

According to the literature of reading comprehension, the average text reading rate for a mature English reader is around 200 to 250 wpm. If an adult individual reads from a computer monitor, it is estimated that he spends 20% to 30% more reading time than he does from papers (Bailey, 1999). Ziefle (1998) investigated the effects on reading performance using hard copy and two resolutions of monitors: 1664x1200 pixels (120 dpi) vs. 832 x 600 pixels (60 dpi). His study found that reading from hard copy was reliably faster (200 wpm versus 180 wpm on screen). In this case, the reading speed range for an adult English reader on a computer monitor would be estimated as 180 to 230 wpm.

[1] The *UDHR in Unicode* database demonstrates the use of Unicode for a wide variety of languages, using the Universal Declaration of Human Rights (UDHR) as a representative text. http://blogamundo.net/lab/wordlengths/The UDHR was selected because it is available in a large number of languages from the Office of the United Nations High Commissioner for Human Rights (OHCHR) at http://www.unhchr.ch/udhr/.

If this reading rate is borrowed as the indicator of a normal reading speed in this study, some raters' reading rates may raise eye-brows. There are three raters, 1, 4 and 7, whose reading rates hit over 300 wpm and their maximum reading rates were even faster than 400 wpm. At such a fast reading speed, raters' text comprehension may suffer significantly. For rater 4 and 7, their standard deviations of reading rate were the highest two among all raters, which indicates that their reading rates varied substantially due to different text features or essay qualities. Rater 1, however, had a remarkably high reading rate across all essays and a medium standard deviation, suggesting that he consistently read faster than other raters.

These different reading behaviors might be accounted for by the individual difference of rater's reading ability. In this study, however, this possibility can be excluded as all of these participants are fluent English readers whose GRE verbal scores are ranked above 70% of their peers. Those non-native speaking participants had obtained a TOEFL score over 627 (paper-pencil test) and they had already studied in a master program for around two years. If rater's reading ability is not taken into consideration, another explanation to this result is that some raters, such as rater 1, were speed reading during their essay grading, suggesting that they might skim, scan or skip some passages. Such a reading behavior, however, may impede their essay comprehension and hence challenge the validity of their scoring.

Studies of speed reading suggest that comprehension declines as a reader increases reading speed above the normal rate. Just and Carpenter (1987) compared the reading comprehension of speed readers and normal readers and found that the normal readers got an overall better understanding of the reading passage. They reported that the speed readers did as well as the normal readers on the general gist of the text, but were worse at details. In fact, the speed readers performed only slightly better than a group of people who simply skimmed

through the passage. In the context of essay grading, as readers must fully comprehend the content of students' writing before assigning essay scores, speed reading may in fact jeopardize the validity and/or reliability of their scoring. In other words, the fact that raters assign an essay score without thorough comprehension of the text determines that no accurate and consistent inferences of the target criterion could be made based on test score. In this study, the reliability of rater 1 and his impact on test validity were further analyzed through other scoring behaviors such as his text reading pattern and his scoring focus.

In addition to raters' reading time, their overall reading patterns were estimated in this study. The visual representations of their linear reading pattern are presented in Figure 4.1, 4.2, 4.3 and 4.4. In these scatter plot charts, the black dots stand for readers' mouse clicks when highlighting sentences during their text reading. The location of the black dots carries both temporal and spatial information about when and where in a text raters made the mouse-click. The X-axis in these charts represents reading time and the Y-axis stands for the length of an essay. Both of these two variables are normalized so that one unit change of time is corresponding to one unit change of essay length.

This two-dimensional chart then depicts the temporal and spatial representations of raters' sentence selection/highlighting during reading, which reflect the overall pattern of raters' text reading. If a rater reads essays at a uniform rate, his overall reading pattern is predicted as a 45-degree linear representation starting from the origin. This linear reading pattern suggests that one unit of his reading time is corresponding to one unit of the total length of essays. The slope of this linear trend stands for the reading speed while the dispersion of these mouse-click dots along the linear pattern represents the degree of changes of a rater's reading rate. The larger the dispersion of these black dots in these charts, the more frequently raters change their reading

speeds due to different text features or essay qualities. If the slope of a linear reading pattern is larger than 45 degrees or if most of the black dots cluster towards the upper range of this chart, this rater's reading rate is overall steady yet faster than the "robot-like" reading rate as he reads more than one unit total length of essays within one unit of his normalized reading time. If the slope of the linear reading pattern is smaller than 45 degrees or if most of the black dots cluster towards the lower part of the chart, this rater's reading rate is slower than the uniform reading rate. In this study, raters had to keep highlighting sentences in order to read essays on the interface. The time and location of their mouse clicks, therefore, were automatically monitored by the rating interface and future processed by the Python-analyzer to estimate raters' reading patterns. The current results report that participants have four major reading patterns that can be illustrated in the following charts.

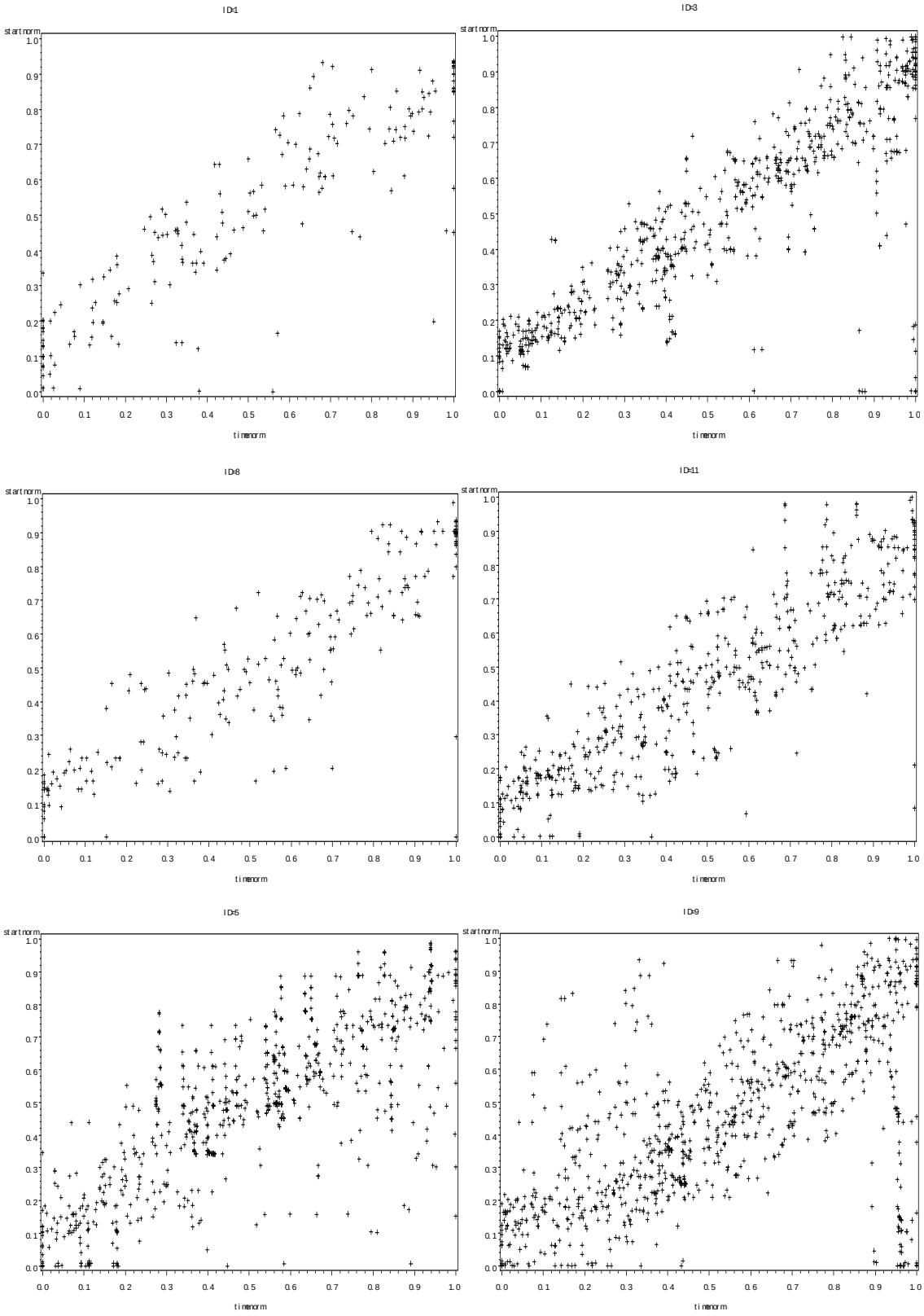


Figure 4.1: The linear reading patterns of reader 1, 3, 8, 9, 5 and 11 (clockwise).

The evident linear patterns in Figure 4.1 demonstrate that these six raters had a linear reading pattern during their essay grading, which suggests that they all had a relatively smooth and consistent reading rate. The fact that the mouse-click dots form one linear line starting from the origin in each chart implies that each rater started reading an essay from the beginning of their reading time and arrived at the end of the essay when the reading time was up. This monolinear reading pattern hence suggests that these raters read each essay for one time only before they reached their scoring decision. The mouse-click dots of rater 1, 3, and 8 cluster around 45 degree line, which indicates that these three did not make frequent reading digressions[2] during their essay grading. The other three raters in Figure 4.1, on the other hand, made more reading regression to previous sentences (shown by dots below the line) or reading projections to the following sentences (shown by dots above the line). This explains why their mouse-click dots have a larger dispersion around the 45-degree linear reading pattern.

The reading patterns of rater 1 and rater 9 demonstrate quite unusual reading behaviors compared to the other four raters in Figure 4.1. The linear line of rater 1's reading pattern suggests a fast reading rate as most of his mouse-click dots cluster above the 45-degree linear trend. This result confirms previous findings of raters' text reading speed. Based on the visual representation of rater 1's text reading pattern, it is plausible to conclude that this rater read each essay at a consistent fast speed. He made only a few reading digressions during text reading, which implies that he did not make frequent comprehension check when grading a sample essay. Quite on the opposite of rater 1, rater 9 made more distant reading digressions as displayed in Figure 4.1. Besides the fact that in general he read most essays for one time, rater 9 tended to skip or skim some sentences in the first half of each text and quite often skimmed the whole

[2] Reading digression refers to a temporary eye-movement departure from the current sentence/phase to the previous/following or a more distant string before the reading of the current subject is resumed.

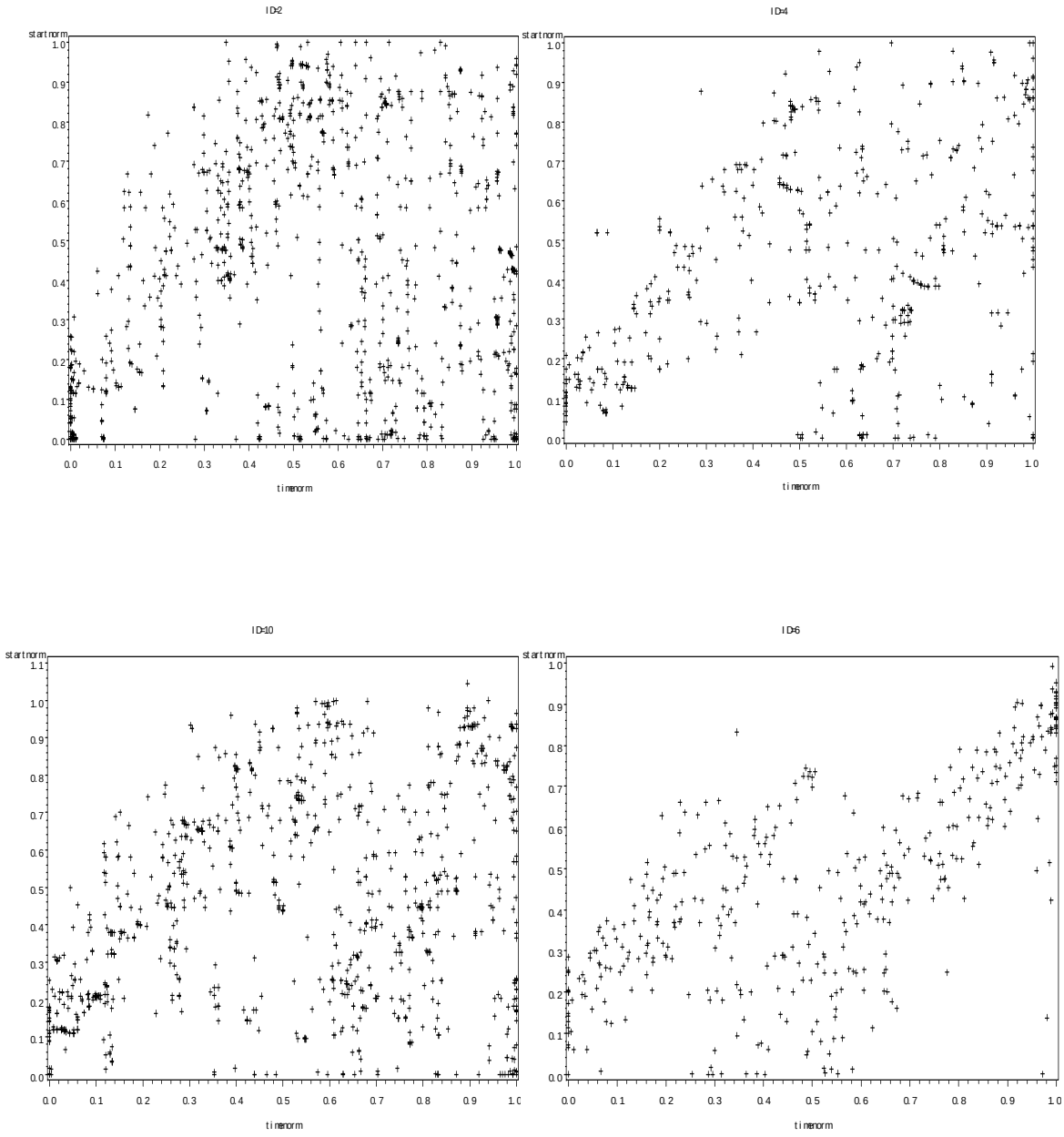
passage again towards the end of his reading. The substantial amount of reading digressions slowed down his reading speed. The fact that most of his mouse-click dots sit below the 45-degree diagonal line infers a low reading rate. This finding is also supported by the results in Table 4.2 where rater 9 is ranked the third slowest reader among twelve.

Compared to the six raters in Figure 4.1, the following raters share a different reading pattern in Figure 4.2. The linear reading trait of these four raters can be represented by two lines that are roughly parallel. The presence of two linear reading patterns provides strong evidences that these four raters read most essays two times. The facts that the upper line is steeper than 45 degree and the lower line starts from the middle of the X-axis suggest that these raters first skimmed the passage at a fast reading rate and then started re-reading the essay from almost the very beginning of the text since the initial point of the lower line is very close to the X-axis. As both of these two lines have a slope larger than 45 degree, raters seemed to read faster than they would normally do if they read each essay once only. Their reading digressions, as we can see from this chart, are much more frequent than that of the first group as the mouse-click dots spread in a larger range.

These raters' frequent reading digressions and their repeated reading suggest a more engaged reading process and a positive impact on their text comprehension. As we've reviewed in previous chapters, text comprehension requires a complex process. Besides the text-based word recognition and syntactic parsing of a sentence, reader must also construct a meaning representation that is coherent at both local and global levels. This process requires readers to determine, for example, what entities pronouns and definite descriptions refer to, and make inferences about relationships between events and entities (Staub and Rayner, 2006). This process also increases the probability of reading regression or digression during the silent

reading of long passages. In this case, given the similar reading ability, readers who repeated reading and had more reading digressions made more efforts to process the text-base information and hence inferred a coherent meaning representation of the reading passage.

This impact of repeated reading behaviors on reading rate and text comprehension has been examined in the psychology of reading. In some short-term experiments, repeated reading was found to yield improved comprehension of the particular passage that was read . Faulkner and Levy (1999) used repeated reading with readers across skill levels and proposed that the benefits of repeated reading for low-skilled readers may be limited to word-level skills, whereas higher skilled readers would improve in reading comprehension as well as rate. Therrien (2004) conducted a meta-analysis to examine the prospective gains of fluency and comprehension as a result of repeated reading. His analysis indicates that repeated reading increases reading fluency and comprehension and can be used as an intervention to increase overall fluency and comprehension ability.



4.2: The linear patterns of rater 2, 4, 6, and 10 (clockwise).

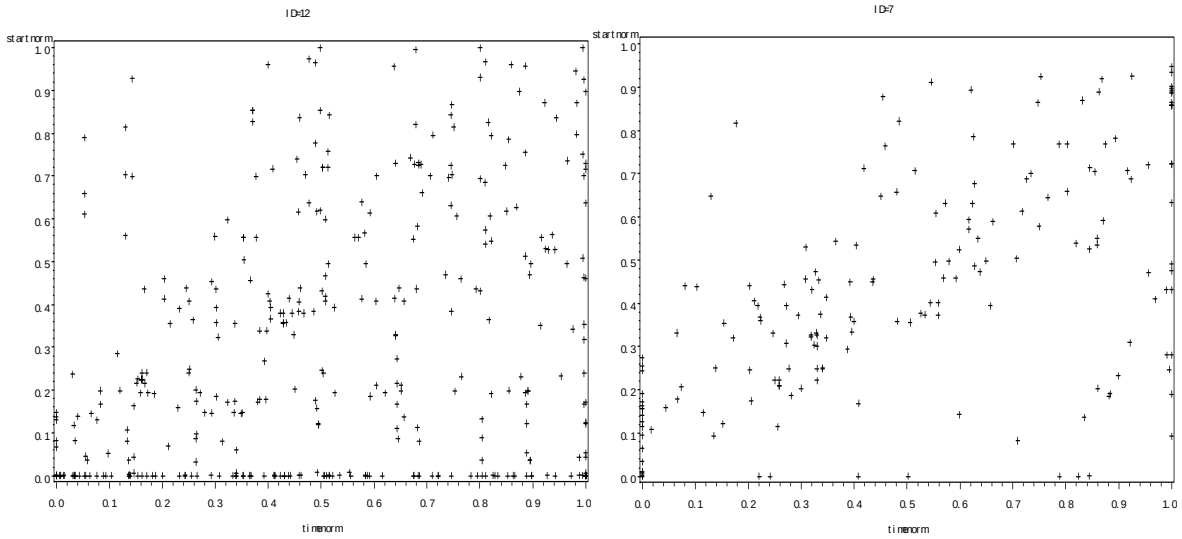


Figure 4.3: The reading patterns of reader 12 and 7 (right).

The reading pattern of the third group of readers/raters, as shown in Figure 4.3, does not demonstrate a clear linear trait. It seems that these two raters constantly make reading regressions or projections, especially rater 12. This figure shows that he skimmed the whole passage a couple of times during reading and his reading frequently regressed to the very beginning or the introduction of an essay.

Raters' different reading traits reported in Figure 4.1-4.3 may be affected by their scoring experience. Among the current participants, seven of them are experienced raters who participated in the EPT rater training and had also scored in operational EPT sessions for over two semesters. Compared to these experienced raters, the other five raters, rater 5, 6, 8, 9 and 11, hadn’t obtained either operational rater training or EPT grading experiences by the time of data collection in this study. However, they were quite familiar with the scoring rubrics of the EPT as they used the same benchmark to evaluate their students’ essays in ESL writing courses for over two semesters. Despite the familiarity of EPT rating benchmarks and ESL essays written by the

same student population, the non-experienced raters had slightly different reading behaviors. Compared to the operational EPT raters, all untrained raters, except for rater 6, demonstrated a monolinear reading pattern and made less reading digressions. The experienced raters, on the other hand, had more diverse reading patterns.

Table 4.3: Estimates of raters' reading pattern: the regression R-square and raters' lps reading rate.

ID	R square	Reading Speed
1	0.7569	25.28
2	0.0214	11.76
3	0.7368	20.94
4	0.1485	26.43
5	0.592	8.74
6	0.4533	18.00
7	0.3354	24.93
8	0.7734	21.16
9	0.4117	8.73
10	0.0742	12.37
11	0.782	16.81
12	0.0712	15.80

In this study, raters' reading patterns were further quantified statistically by regressing the normalized length of essay onto the normalized reading time. Table 4.3 provides summary statistics of raters' regression R-square and related reading rates. The larger the R-square, the larger probability that the temporal-spatial representation of a rater's reading pattern regresses towards a linear line and the smaller the reading digression rate, suggesting a less probability that readers regress to previous essay chunks or suddenly shift their attention to the following or more distant strings . This result coincides very well with our previous observations in Figure 4.1-4.3 and these two indicators (regression R-square and reading rate) provide useful information to interpret raters' reading comprehension. First of all, their reading speed is highly correlated with the linearity pattern in Figure 4.1-4.3. Those who had a high reading rate and

high regression R-square, such as rater 1, 3 and 8, demonstrated a clear monolinear reading pattern without many reading digressions. As we've discussed in earlier paragraphs, this finding provides evidences of impaired text comprehension. Since these raters read at unusually high rates and they did not make frequent comprehension check during essay grading, they might not be able to fully comprehend the text base information and/or construct a meaningful global representation of an essay. On the other hand, raters who had a low regression R-square not only made more frequent reading digressions but also demonstrated two-line or non-linear reading patterns. Last but not least, low regression R-squares tend to be associated with experienced raters, such as rater 2, 4, 8 and 12. The first three raters repeated reading each essay during grading and the last one had a non-linear reading trait. All of them had made abundant reading digressions to check their text comprehension during the experiment.

In this study, raters' text comprehension was indirectly addressed through raters' score assignment and rater reliability. As all raters in this study read the same set of essays and they were equally acquainted with the scoring criteria, the reliability of their scoring depends on their text comprehension and their judgment of essay qualities. In this study, it is hypothesized that the reliability of a rater's scoring would be jeopardized if his unusual reading behavior may impair his text comprehension at both text base and discourse levels.

Raters' scoring assignments and the correlation between these holistic scores are reported in Table 4.4 and 4.5. Despite the fact that the standard deviation of rater 1's scoring assignment is the largest among raters, the results in Table 4.5 show no significant difference between score means. The results reported in Table 4.5, however, demonstrate that some raters' scoring judgments are not statistically correlated with the scores assigned by others. For example, the essay scores assigned by rater 1 were not significantly correlated with that of seven other raters

and also a comparatively low inter-rater reliability with the rest four raters. This result indicates a low agreement or concordance between rater 1 and the other raters. This disagreement is strongly associated with raters' different reading behaviors during essay grading. Those raters who had high reading digression rates and comparatively low reading rate, for example, rater 2, 10 and 12 in Table 4.5, have a higher inter-rater reliably. This finding supports the Hypothesis 1 in this study:

Hypothesis 1: A high reading digression rate and a low reading rate indicate an engaged reading comprehension process during essay grading, hence these indices are positively associated with rater reliability in a writing test.

According to the results in Table 4.5 and 4.3, rater's reading digression rate itself is associated with their score agreement; therefore, it could be viewed as an indicator of rater concordance. A high inter-rater reliability is in general associated with a high reading digression rate and vice-versa. Compared to inexperienced raters, most EPT raters who had training and grading experiences (except for rater 1) tend to have a higher reading digression rate and thus have a higher inter-rater reliability. It seems that experienced raters internalized the scoring criteria during their training and previous scoring practice and they knew already what to look for when grading an essay. On the other hand, the fact that raters made frequent reading digressions and repeatedly read an essay also helps them to construct meaningful inferences of the writing discourse, hence enabling them to reach an accurate judgment of essay quality.

Table 4.4: Summery Statistics of Raters' Score Assignment.

Rater ID	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12
Mean	2.45	2.80	2.70	3.15	2.35	2.45	2.59	2.45	2.60	2.50	2.80	2.75
Std	0.94	0.52	0.66	0.75	0.49	0.69	0.51	0.60	0.75	0.61	0.70	0.64
No. Article	19	19	20	18	20	18	20	20	20	20	20	19

Table 4.5: Estimate of Inter-rater Reliability.

Correlation	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12
r1	1.00											
r2	0.64	1.00										
r3	***0.35***	0.64	1.00									
r4	0.45	0.62	0.56	1.00								
r5	***0.22***	0.49	0.54	0.45	1.00							
r6	0.44	0.65	0.74	0.52	0.80	1.00						
r7	0.52	0.66	0.54	***0.26***	***0.46***	0.66	1.00					
r8	***0.29***	0.50	***0.40***	***0.41***	***0.30***	0.53	0.69	1.00				
r9	***0.26***	0.51	0.65	0.40	***0.31***	0.62	0.56	0.59	1.00			
r10	***0.39***	0.61	0.73	0.66	0.62	0.81	0.64	0.52	0.47	1.00		
r11	***0.41***	***0.43***	***0.34***	***0.15***	***0.37***	0.61	0.55	0.51	***0.32***	0.59	1.00	
r12	***0.31***	0.69	0.86	0.65	0.62	0.69	0.54	0.52	0.69	0.67	***0.38***	1.00
# Non-Corr	7	1	3	3	5	0	2	4	4	1	6	2

4.2 Rater's Attention Distribution

For each essay, the average reading rates across essays are demonstrated in Table 4.6, which reports that raters' reading rates vary substantially due to certain essay features. For example, the mean reading rate for essay 9 and 10 are over 22 lps, while that of essay 1 and 2 are around 10 to 13 lps. This result suggests that readers may find it more difficult or easier to read certain essays before they reach their scoring decisions.

Table 4.6: The letter-per-second reading rate for each essay.

Textid	N	Mean	Std	Max	Min
1	12	10.44	5.08	2.49	22.38
2	12	13.71	6.36	3.19	27.37
3	12	15.17	10.08	6.22	42.66
4	11	13.92	7.13	5.53	27.65
5	12	17.12	10.6	7.28	44.07
6	12	16.64	9.79	5.82	38.67
7	12	15.68	6.7	8.99	28.22
8	12	15.89	12.5	5.38	48.07
9	11	23.56	12.76	6.1	49.29
10	12	22.24	12.36	7.19	44.09
11	11	18.48	8.47	5.51	38.17
12	12	15.16	6.42	4.96	26.84
13	12	15.7	8.87	4.06	31.99
14	12	18.3	6.86	8.5	30.38
15	12	18.32	8.43	5.86	31.85
16	10	17.84	7.05	5.88	30.33
17	12	23.07	8.44	9.86	42.37
18	11	19.11	8.74	8.78	42.54
19	12	19.1	9.24	5.67	33.85
20	11	21.31	6.71	13.83	33.16

In this study, no strong correlation is observed between essay score and the mean reading time associated with each essay. The different reading rates across essay can be accounted for by various essay features shown in Table 4.7. Previous reading studies have reported that text reading rate is significantly affected by text features. This finding is supported by correlations between seven essay features and rater's reading rate in Table 4.7.

Table 4.7: Correlations between seven essay features and rater's reading rate.

	vocab	word	sentences	subsent	trancount	trantype	freq
Unit Time	0.12	0.24	0.08	0.24	-0.35	-0.37	0.15

In this table, *vocab* is defined as the total number of vocabulary shown in 20 essays excluding stop words (defined as words that have a high frequency and low semantic information, such as 'the'). *Word* refers to the total number of vocabulary including stop words. *Sentences* stands for the total number of sentence and *subsent* the total number of sub-sentences. *Trancount* means the total number of transitional words and *trantype* the total number of different transitional words in these essays. *Freq* refers to the total word frequency. The word frequency was estimated with Brown Corpus, which mainly consists of newspaper articles. In this study, *Freq* is referred as the weighted average word frequency in essays, with the weight defined as word frequency of the vocabulary from Brown Corpus.

Raters' total reading time for each essay is positively correlated with essay features including *total number of words, total number of sub-sentences* and is negatively correlated with number and type of *transitional words*. The fact that both positive and negative effects on reading time are observed implies that some essay features may facilitate raters' text comprehension and accelerate reader's reading rate, while other features impair their reading comprehension. For example, the number of transitional devices and their logical categories are negatively correlated with reading time, which suggests that raters spend less time on essays with more transitional devices that belong to various logical categories (e.g. causal, temporal and compare/contrast). This result is consistent with the findings in the reading studies of situation models which suggest that the presence of transitional devices help to construct text coherence and thus facilitate readers' integration of upcoming information into the evolving mental representation (Zwaan, et al., 1998). On the other hand, other sentence features, especially the total number of word (*word*) and the number of clause in an essay (*subsent)*, are positively correlated with reading time. Readers spend more time reading longer essays that have a larger

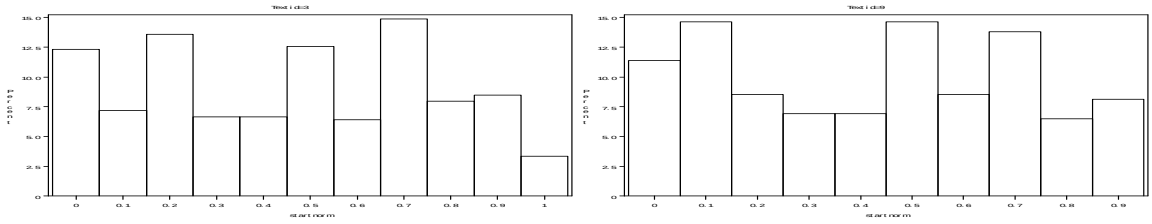
vocabulary variety (*word*, *sentences* and *subsent*) and more complex syntactic structures (*subsent*). This finding also confirms the conclusion from previous studies of eye movement in reading comprehension (e.g. Hyönä and Vainio, 2001; Rayner, 1998).

The positive correlation between word frequency and reading time, however, is contradictory to previous findings that predict a negative effect of word frequency on reading time. In this study, it seems that readers spend more time if the average word frequency in a text is higher. This surprising result may be explained by two experiment conditions: 1) the word frequency might be biased by spelling errors of the low frequency words in this study, thus it was not accurately estimated, 2) within the current test context, the high frequency words had their synonyms in the EPT reading passage or the lecture that raters were quite familiarly with. Therefore, the low frequency words did not impede readers' comprehension.

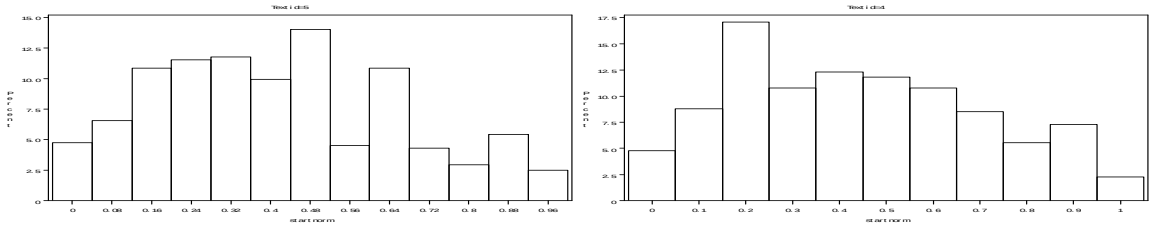
Based on raters' sentence selection/highlighting, the distribution of their attention on each essay were estimated via the distribution of their total reading time on each essay. Evident patterns of raters' attention distribution (measured as time spent on certain parts of an essay) can be observed for seven essays in Figure 4.4, in which X-axis represents the essay length and Y-axis the reading time. In this figure, we can see that some parts of these essays receive more attention as raters spent more time on these chunks. There are four major attention distribution patterns identified in this study: 1) uniform distribution. Raters' reading time is evenly distributed to each sentence in essay 3 and 9. Raters did not pay extra attention to a particular chunk in these two essays. 2) Unimodal distribution. Most raters spent more time reading the body of essay 5 and 4 and skimmed the beginning and ending parts of these two essays. 3) Bimodal distribution. For essay 17 and 16, reader's attention evenly clusters around the two chunks located right after the beginning and before the ending of the text. 4) Trimodal

distribution. Essay 11 draws raters' attention to the introduction, conclusion and the very middle part of the text. These different attention distributions lead us to a plausible conclusion that raters' reading time is affected by the feature and writing quality of a particular essay such as essay organization, content, syntactic complexity and logical coherence. This finding is supported by the literature of reading comprehension. For example, if an essay contains a syntactic anomaly that strongly impedes comprehension, readers are expected to spend more time reading or re-reading this chunk or adjacent scripts as well (Braze et al., 2002; Deutsch and Bentin, 2001; Ni et al., 1998; Pearlmutter et al., 1999). In this case, reading time can be viewed as a robust indicator of reader's attention distribution as we observe in Figure 4.4.

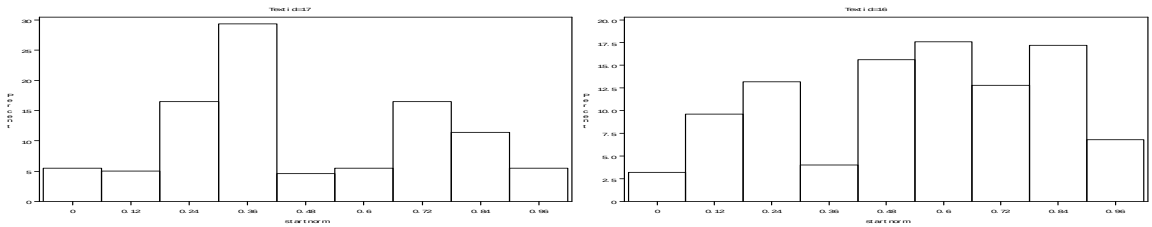
Uniform



Unimodal



Bimodal



Trimodal

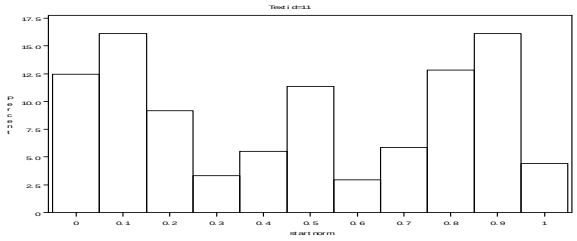
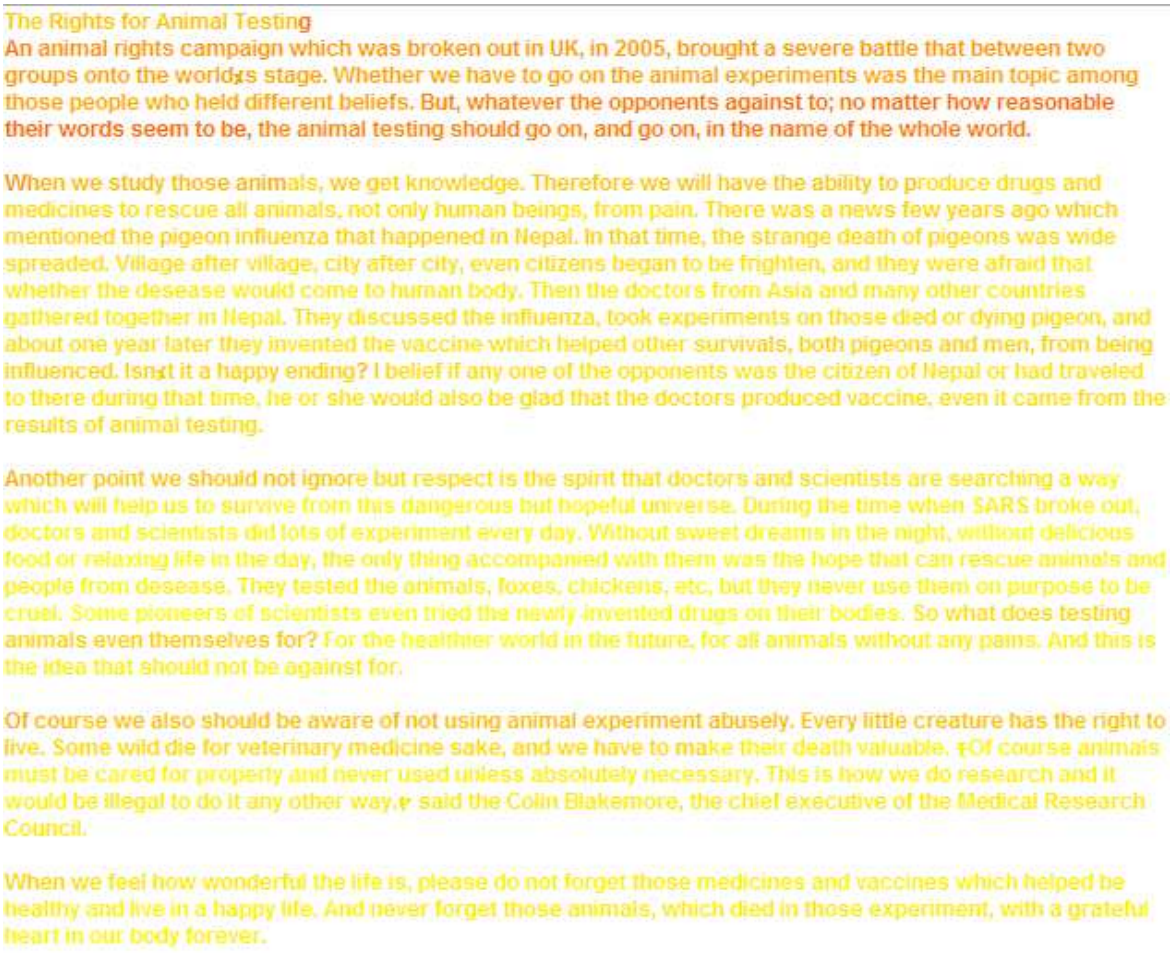


Figure 4.4: The distribution of raters' attention across essays.

The fact that raters have a common attention spread on a particular essay may signal the existence of a shared reading pattern among raters, which may provide behavioral evidence to evaluate the validity and reliability of a writing test. If a rater did not distribute his attention the way other raters did on a given essay, it is highly probable that this rater was lack of attention during text reading or he paid more attention to irrelevant response categories that should not have been focused on. As a result, this rater might not able to assign a score from the shared/pre-designed scoring criteria, thus reducing the rater agreement and test reliability. Another sequential problem is that the test score fails to represent or represents less precisely test takers' ability level for the target construct as this rater may evaluate an essay based on a construct-irrelevant variability. In this case, a threat to test validity can be predicted as well.

Figure 4.4 depicts a rough distribution of reading attention among different parts of an essay. As an alternate method to display raters' attention, a text-base representation of their reading time demonstrates more detailed textual information of the strings or chunks that raters focus on. By visualizing the attention "hot spot" (defined as sentences/phrases that receive more attention) on each essay, we are able to directly look at the text chunks that cost readers more time to read and hence analyze their features. In the hotspot attention display, for a certain area in an essay, the color goes from yellow to red as its related reading time increases. That is to say, the darker the scripts, the more attention these scripts have obtained from all readers. For

example, raters' attention distribution is represented by different font colors on essay 11 and 5 in Figure 4.5 and 4.6, respectively.


The Rights for Animal Testing

An animal rights campaign which was broken out in UK, in 2005, brought a severe battle that between two groups onto the worldɀs stage. Whether we have to go on the animal experiments was the main topic among those people who held different beliefs. But, whatever the opponents against to; no matter how reasonable their words seem to be, the animal testing should go on, and go on, in the name of the whole world.

When we study those animals, we get knowledge. Therefore we will have the ability to produce drugs and medicines to rescue all animals, not only human beings, from pain. There was a news few years ago which mentioned the pigeon influenza that happened in Nepal. In that time, the strange death of pigeons was wide spreaded. Village after village, city after city, even citizens began to be frighten, and they were afraid that whether the desease would come to human body. Then the doctors from Asia and many other countries gathered together in Nepal. They discussed the influenza, took experiments on those died or dying pigeon, and about one year later they invented the vaccine which helped other survivals, both pigeons and men, from being influenced. Isnɀt it a happy ending? I belief if any one of the opponents was the citizen of Nepal or had traveled to there during that time, he or she would also be glad that the doctors produced vaccine, even it came from the results of animal testing.

Another point we should not ignore but respect is the spirit that doctors and scientists are searching a way which will help us to survive from this dangerous but hopeful universe. During the time when SARS broke out, doctors and scientists did lots of experiment every day. Without sweet dreams in the night, without delicious food or relaxing life in the day, the only thing accompanied with them was the hope that can rescue animals and people from desease. They tested the animals, foxes, chickens, etc, but they never use them on purpose to be cruel. Some pioneers of scientists even tried the newly invented drugs on their bodies. So what does testing animals even themselves for? For the healthier world in the future, for all animals without any pains. And this is the idea that should not be against for.

Of course we also should be aware of not using animal experiment abusely. Every little creature has the right to live. Some wild die for veterinary medicine sake, and we have to make their death valuable. ɁOf course animals must be cared for properly and never used unless absolutely necessary. This is how we do research and it would be illegal to do it any other way.ɂ said the Colin Blakemore, the chief executive of the Medical Research Council.

When we feel how wonderful the life is, please do not forget those medicines and vaccines which helped be healthy and live in a happy life. And never forget those animals, which died in those experiment, with a grateful heart in our body forever.

Figure 4.5: The hotspot display of raters' reading attention for essay 11.

The hotspots of raters' attention are associated with certain text features. In Figure 4.4, essay 11 is a text with a uniform distribution of readers' attention as there is no significant attention cluster observed from the histogram chart. If we look at Figure 4.5, however, some attention hotspots are identified as readers spent relatively more time reading the thesis statement, the topic sentences in each body paragraph and the transitional devices in this text. Similar features of the hot spots are also observed in Figure 4.6, in which the hot-spot trait is

consistent with the unimodal distribution in Figure 4.4. Readers seem to pay most attention to the body part of this essay. If we take a closer look at the color of sentences, we can find that most attention hotspots cluster around the following strings: 1) Thesis statement and adjacent chunks. The red color of the second paragraph indicates that most raters spent more time reading this paragraph as the thesis statement of this essay sits in this paragraph. 2) Topic sentence. The first sentences in paragraph 3 to 6, as the topic sentence, show a relatively darker color, suggesting that these sentences receive more attention from raters. 3) Sentences carrying transitional devices. Among those "hot" sentences, a large variety of sentence connectors are observed. For example, *therefore* in the last sentence of paragraph 2, and *first of all*, *second*, *third* and *in summary* at the beginning of paragraph 3 to 6. Readers in general spent relatively more reading time on the second paragraph, but they paid even more attention to sentence connectors, e.g. *according to, thus* and *besides*.

Figure 4.6: The hotspot display of raters' reading attention for essay

These findings again are supported by reading studies that focus on essay responses and text coherence. The fact that raters spent more time reading topic sentences and thesis statement can by interpreted by previous findings about raters' response to different essay qualities. As raters base their judgments primarily on the content and organization of student writing, essay chunks (e.g. thesis statement or topic sentence) that are categorized into these two criteria are expected to attract more attention (Freedman 1979, 1981, 1984; Freedman & Calfee, 1983). On the other hand, readers' attention on transitional devices confirm linguists' claims that connectives help to construct a coherent text representation and the presence of sentence

connectors is positively correlated with readers' text comprehension (Van den Broek, 1988; Van den Broek, et al., 2001). In this study, raters' attention on transitional devices is predicted due to the fact that sentence connectors help to construct logical coherence for the development of an argumentative essay. As raters pay more attention to these response criteria, they would naturally search for sentence connectors as evidences of text coherence.

Besides raters' reading time, their score assignment is also strongly correlated with certain essay features. Table 4.8 reports the correlation between eleven essay features and the scores assigned by twelve raters. In this table, *Word* stands for the total number of words in each essay including repeated words and stop words. *Vocabulary* refers to the total number of non-repeated words excluding stop words. *Sentence Length* is defined as the total letters in a sentence. *Sentences* stands for the total number of sentence and *Subsentences* the total number of sub-sentences. *Category of Tran. Word* indicates the types of transitional words and *Tran. Word* the total number of transitional words. *Essay length* is estimated through total letters and punctuations in an essay and *Word Length* the average number of letters in a word. *Word Per. Sentence* stands for the average number of words in a sentence for each essay. *Word Frequency* refers to the weighted average word frequency in each essay, with the weight defined as word frequency of the vocabulary from Brown Corpus.

Table 4.8: Correlation between Essay Features and Essay Scores

Essay Features	mean	min	max	std	Corr. Coefficient	P-value
Word	402.55	295.00	533.00	74.21	0.28	0.23
Vocabulary	***125.55***	***83.00***	***171.00***	***25.02***	***0.47***	***0.03***
Sentence Length	***89.15***	***57.16***	***117.48***	***14.96***	***0.49***	***0.03***
Sentence	27.75	20.00	38.00	4.53	-0.13	0.59
Subsentences	51.40	27.00	68.00	10.73	0.25	0.29
Category of Tran. Word	13.55	7.00	24.00	5.10	0.16	0.50
Tran. Word	8.70	4.00	13.00	3.21	0.34	0.14
Essay Length	*2444.45*	*1783.00*	*3172.00*	*424.70*	*0.37*	*0.10*
Word Length	6.09	5.61	6.78	0.26	0.25	0.28
Word Per. Sentence	*14.68*	*9.32*	*19.74*	*2.56*	*0.39*	*0.09*
Word Frequency	0.02	0.01	0.04	0.01	-0.09	0.71

According to Table 4.8, certain essay features such as *Vocabulary* and *Sentence Length* are significantly correlated with essay scores. This result implies that if an essay contains more non-repeated words and long sentences (generally speaking a sentence with a more sophisticated structure), it tends to obtain a higher essay score. Besides these two indicators, *Essay Length* and *Word Per Sentence* also demonstrate a relatively high correlation with individual essay score. These findings confirm the interaction between raters and texts, hence supporting the second hypothesis.

Hypothesis 2: If there is an interaction between rater and essay writer, raters' scoring decision is associated with essay features.

In this study, both raters' reading time and their scoring decision making are affected by linguistic features, e.g. characteristics of vocabulary and sentence, in an essay. The current results imply that if a text contains long sentences composed of sub-phrase and a large number of non-repeated vocabularies, raters would spend more time reading this passage and tend to leave a relatively high score. This is a valid prediction based on the findings in the literature of

automated essay scoring (e.g. Burstein, et. al. 1998; Valenti, et. al., 2003). As one of the earliest implementation of automated essay grading engine, Project Essay Grade (PEG) primarily relies on style analysis of surface linguistic features of a text. Therefore, an essay is predominantly graded based on prescribed writing "proxes". Among these "proxes", essay length defined as the amount of words in an essay is viewed as the presentation of writing fluency and word length as the indication of diction as less common words are often longer (Valenti, et. al., 2003). Besides essay length and word length, the size of vocabulary is also used as a robust feature that reflects writing qualities (Burstein, et. al, 1998).

4.3 Rater's Decision Making

4.3.1 The Dynamic Information: Verbatim Annotation and Score Comments

In this study, raters were required to annotate sentences/phrases from sample essay as the evidence of their score decision. They were also asked to leave comments and answer rating questions on the interface. Raters' online scoring events including annotating and commenting were hence automatically monitored and analyzed by the IRE.

Table 4.9 demonstrates the summary statistics of raters' scoring comments, which could be divided into two major categories: positive comments that acknowledge writer's strength or negative comments that point out the flaws in an essay. If a comment contains both positive and negative essay features, it will be counted in the category of *"both"*. Table 4.9 displays the type of comment assorted by rater ID. Individual differences regarding raters' commenting preference are observed in this table: the proportion of positive comments versus negative comments varies across raters. However, generally speaking, raters left more negative comments than positive

ones. This overall pattern of raters' commenting preference suggests that it may be easier for raters to identify the ill-formed essay features when evaluating essay qualities.

Table 4.9: Summary statistics of raters' comment type.

ID	Negative	both	Positive	GrandTota	Good/Bad
1	12		3	15	0.25
2	23		4	27	0.173913
3	39	3	11	53	0.282051
4	17	2	17	36	1
5	53		12	65	0.226415
6	41	16	27	84	0.658537
7	30	5	11	46	0.366667
8	59	13	21	93	0.355932
9	52		6	58	0.115385
10	34	3	3	40	0.088235
11	51	4	9	64	0.176471
12	49	5	1	55	0.020408
Grand Total	460	51	125	636	0.271739

The categorization of raters' scoring comments is also affected by essay features and writing qualities. Table 4.10 displays the total number of positive, negative and neutral comments sorted by essay. The ratio of positive vs negative comments is highly correlated with essay scores. For a well-structure text, such as essay 17, the total amount of positive comments outweighed that of the negative ones; hence the related positive/negative comment ratio is one of the largest in Table 4.10. On the other hand, if an essay is ill-written, e.g. text 4, raters tend to focus on the imperfections of this text and leave negative critiques.

Table 4.10: Summary statistics of raters' comment type.

textid	bad	both	good	Grand Total	Good/Bad
1	22	6	6	34	0.27
2	24		12	36	0.50
3	20	2	13	35	0.65
4	33		4	37	0.12
5	17	2	9	28	0.53
6	35	2	1	38	0.03
7	27	4	3	34	0.11
8	29	3	2	34	0.07
9	20	1	1	22	0.05
10	15	3	9	27	0.60
11	22	7	5	34	0.23
12	27	1	4	32	0.15
13	39	1	5	45	0.13
14	10	3	15	28	1.50
15	34	1	1	36	0.03
16	24	2	3	29	0.13
17	9	4	15	28	1.67
18	26	2	5	33	0.19
19	20	4	2	26	0.10
20	7	3	10	20	1.43
Grand Total	460	51	125	636	0.27

During raters' essay grading, they also made either positive or negative verbatim annotations as the evidence of their scoring judgment. A strong correlation is observed between the proportion of positive/negative annotations and the average score for each essay. To sum up, results from Table 4.10 and 4.11 suggest that the experiment essays are associated with different ratios of positive versus negative comments and annotations. These two ratios are significantly correlated with the average essay score (F=0.763 for comment, $p < 0.001$; F=0.752 for annotation, $p < 0.001$). The strong correlation between raters' score assignment and their scoring behaviours suggest that the way they make their decision is reflected not only in their score assignment but also in their scoring behaviours such as sentence selection, verbatim annotation

and comments, as these behaviors are inseparable part of their rating process. This conclusion confirms the Hypothesis 3 proposed in the current investigation.

Table 4.11: Number of positive and negative annotation by essay

textid	BAD	GOOD	Grand Total	Good/BAD
1	25	24	49	0.96
2	39	24	63	0.62
3	23	35	58	1.52
4	48	18	66	0.38
5	36	29	65	0.81
6	45	5	50	0.11
7	31	13	44	0.42
8	27	11	38	0.41
9	25	11	36	0.44
10	21	21	42	1.00
11	27	11	38	0.41
12	35	12	47	0.34
13	36	19	55	0.53
14	15	30	45	2.00
15	55	2	57	0.04
16	23	11	34	0.48
17	7	25	32	3.57
18	27	26	53	0.96
19	21	18	39	0.86
20	8	36	44	4.50
Grand Total	574	381	955	0.66

An alternate method to demonstrate raters' scoring foci on essay features and response criteria is to map on the text the verbatim annotations and scoring comments that raters made as the online evidence of their scoring decision making. In Figure 4.7, the verbatim annotations for an ill-formed essay and a well-written one are displayed. The blue font represents the negative annotations made by all raters, the warm color stands for the positive annotations, and the black scripts mark the location where raters inserted their scoring comments. The darker the color is for a certain text chunk, the more frequent that readers annotated it as scoring evidence during grading. As Figure 4.7 shows, the quality of the upper essay is quite low (received a score of 2),

most annotations are associated with negative essay features. For the lower essay, however, most comments emphasize in the strengths of this text. There seems to be no remarkable pattern of the distribution of comment insertion based on this figure.

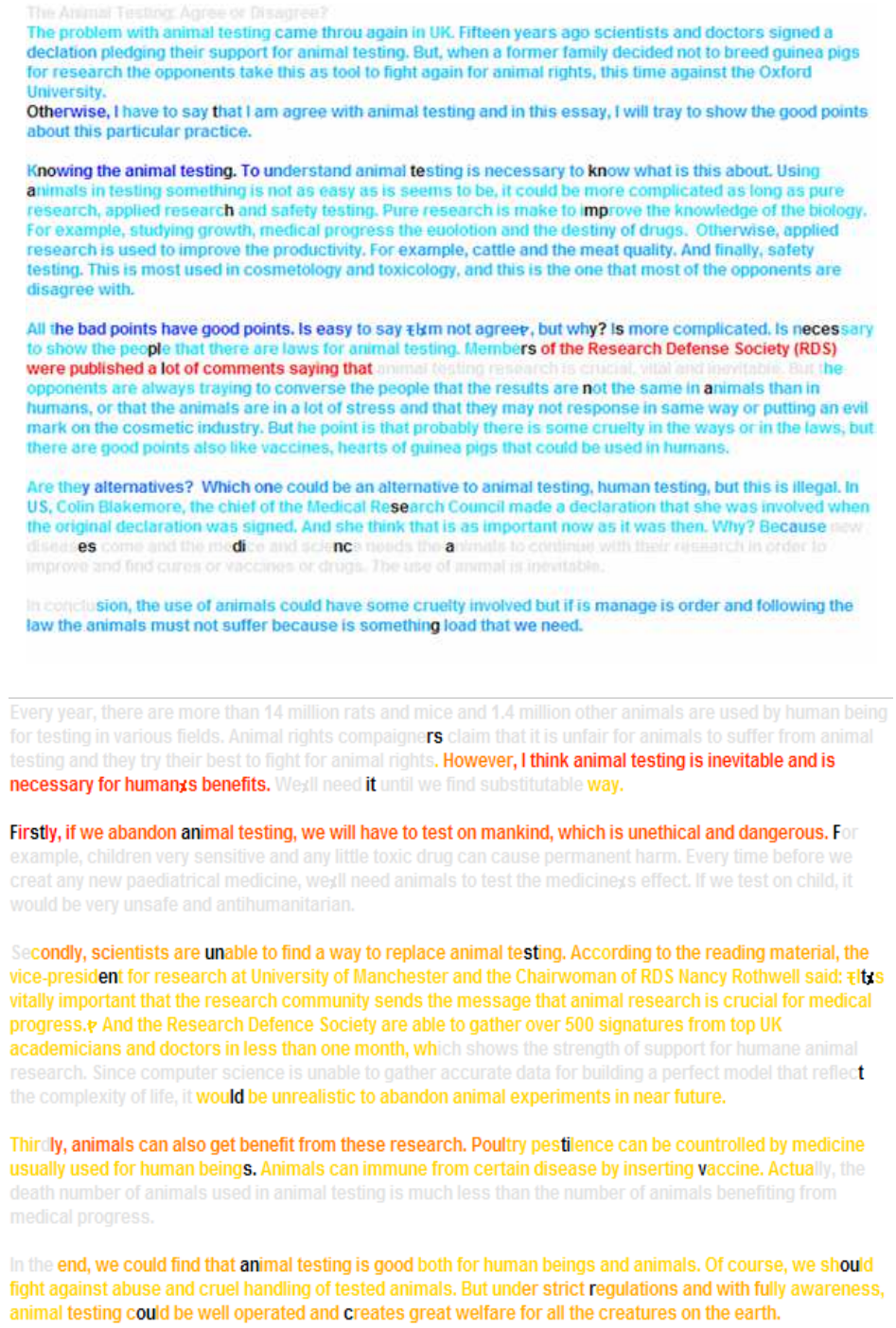


Figure 4.7: Hotspots of verbatim annotation for an ill-formed essay (upper) and a well written essay.

Figure 4.8: Hotspots of verbatim annotations for essay 5.

In Figure 4.8, the distribution of positive and negative annotations is more balanced through text. The location of these verbatim annotations and the feature of adjacent phrases in this figure indicate that, when making annotations, raters focus on certain shared scoring criteria such as idea development, organization and documentation skills. For example, the positive hot spots in Figure 4.8 sit on the topic sentence of each paragraph, which can be viewed as evidences of well-structured development of argumentation. Another positive hot spot is the phrase that indicates writer's appropriate documentation of source materials provided in the writing test. As the color of this phrase is dark in red, many raters have noticed this documentation evidence and

have selected this sentence as positive scoring evidence. This result indicates that raters follow a shared scoring criterion defined as *plagiarism* in the rating rubric. The negative annotations are more difficult to categorize into a certain scoring dimension. By looking at the content of annotated sentences and their location, however, we can tell that the negative annotations are more of the essay content level, putting more emphasis in essay development than in overall text organization.

Raters' focus on the criterion of essay development is also reflected in their scoring comments which have been categorized into five scoring dimensions including 1) text organization, 2) essay development, 3) grammar and lexical choice, 4) plagiarism and 5) extra-rubric qualities such as writing skills and rhetorical strategies. Table 4.12 provides summary statistics of the criteria that raters commented on. In general, raters' comments are mostly associated with *essay development*, followed by *grammar, plagiarism, extra-rubric qualities* and *essay organization*.

Table 4.12: Summary of raters' essay comment type.

Rater ID	1	2	3	4	5	Grand Total
1	2	3	4	5	1	15
2	3	17	0	5	2	27
3	0	39	4	8	2	53
4	1	20	7	4	5	36
5	0	50	13	2	1	65
6	4	51	20	12	3	84
7	2	30	6	6	3	46
8	1	70	11	23	0	93
9	1	33	17	4	4	58
10	0	33	5	0	3	40
11	2	48	15	3	0	64
12	1	35	9	2	8	55
Grand Total	17	429	111	74	32	636

It seems that most raters viewed *essay development* the most fundamental scoring dimension during their essay grading. However, the order of importance among these five scoring dimensions are quite contradictory to the instructions that raters received in the training session. Before the data collection in the present study, a 60-minute training session was delivered to all participants. Each rater was given a copy of the complete EPT scoring benchmarks where five scoring dimensions and relative performance evidences were listed. After reviewing the scoring rubrics individually, raters were assigned to grade four sample essays that represent four scale levels of EPT writing. A set of recalibration answer keys was given to raters after their grading so that they could compare the grades they assigned with the standard placement results. A short group discussion was held after the placement check to help raters to discuss with their peers the weight of each scoring dimension during essay grading and how to distinguish essays placements that are of two adjacent scale levels. During the discussion, raters were instructed to pay most attention to the scoring dimensions of *text organization* and *essay development.* Raters were specifically informed that they should not focus on students' grammatical errors unless it impedes their text comprehension. Based on the instruction of rater training/recalibration, the most important scoring aspect is *text organization,* followed by *idea development, plagiarism* and *grammar and lexical choice*.

One possible explanation to this discrepancy is that most essays had already displayed a clear organization as the writing prompt required test takers to produce an argumentative text with a clear introduction, body and conclusion. Therefore, it might be less necessary for raters to comment on this criterion. In addition, it may be easier for raters to provide comments on the surface structures of an essay rather than to critique essay organization at a global level.

Different scoring foci were also observed between trained and untrained raters. In Table 4.12, the scoring criterion of *grammar and lexical choice* is viewed the second most important scoring dimension. The remarkable amount of comments on this dimension is contradictory to the content of the EPT rater training, in which raters were explicitly instructed that the focus of the EPT test is not students' grammar knowledge but their academic writing ability in producing an argumentative essay. If raters' comment type accurately reflects their scoring emphasis, this discrepancy between test construct, rater training and rating criteria may jeopardize test validity. Fortunately, there were only five raters whose scoring comments were closely related to grammatical features: rater 5, 6, 8, 9, and11. All of these raters were relatively new ESL TAs who had not been trained to grade operational EPT essays by the time of data collection. The lack of EPT grading experience explains their attention to grammatical and lexical features in EPT essays. The fact that untrained raters tend to over emphasize the importance of grammar and lexical choice provides useful information for the modification of rater training.

4.3.2 The Static Information: Post rating questions and Essay scores

Besides the dynamic data that recorded raters' moment-to-moment decision making, self-reported rater responses were also collected from the post-essay questions. Raters were asked to answer four questions after grading each essay. The first two were multiple choice questions, asking raters which scoring criteria that they paid most or least attention when grading an essay. The next two short-answer questions required them to specify the strengths and weaknesses of every essay. Raters' answers to two multiple choice questions are reported in Table 4.13 and 4.14.

Table 4.13: Summary of the scores that are involved in raters' response to short answer questions.

ID	1	2	3	4	Grand Total
1	8	7	19	6	40
2	10	12	13	5	40
3	1	18	2	19	40
4	12	9	6	13	40
5	19	1		20	40
6	7	14	19		40
7	2	18	4	16	40
8	3	16	19	2	40
9	9	10	8	11	38
10	5	14	7	14	40
11	2	18	1	19	40
12	2	17	19	2	40
Grand Total	80	154	117	127	478

Table 4.13 shows twelve raters' score choices of four post rating questions. If we compare the results of Table 4.13 and 4.12, a discrepancy between raters' self reported thoughts and their online scoring behaviors can be observed. For example, many raters, such as rater 1, 2, 4, and 5, self-reported that they believed text organization is the most important aspect to evaluate sample essays; while raters' total counts of their scoring comments in this dimension suggest otherwise. Many of them totally overlook this essay criterion when they left critiques. In fact, *text organization* attracted the least attention among raters. According to Table 4.13, rater 1, 6 and 8 all reported that the role of *grammar and lexical choice* should not be overemphasized during essay grading as they ranked it as the least important scoring dimension. Their scoring comments, however, demonstrate a strong tendency that these raters searched for grammatical errors when reading essays as they left quite a large amount of grammar-related comments. These results infer that raters' self-reported data are not always consistent with their actual scoring behaviors. This finding implies that the current experiment instrument may provide supplementary information of raters' decision making process for related survey studies since

raters' retrospective report may not be the accurate reflection of what they think and/or what they do.

Table 4.14 demonstrates that most raters view *idea development* the most important scoring dimension and *text organization* the second most important dimension. When they were asked what scoring is the least important among the four listed in the rating rubrics, most raters chose *plagiarism* rather than *grammar and lexical choice.* These results confirm raters' perception of the ranking of four scoring aspects from their online grading behaviors. The self-reported data provides similar focus when raters made their score judgment as it is demonstrated by rater annotations and comments. Twelve essay raters ranked the importance of four rating dimensions from *development* as the highest followed by *plagiarism, grammar* and *organization* the lowest. Despite the fact that essay organization was underrepresented in essay rating, there was a consensus among raters about what scoring criteria they took into consideration and how important these criteria were to determine the final essay scores.

Table 4.14: Raters' responses to two multiple choice questions.

answerid	1	2	3	4	Grand Total
1	75	151	6	7	242
2	5	3	111	120	242
Grand Total	80	154	117	127	484

Notes: The row ID stands for the first two multiple choice questions and the column ID refers to four scoring dimensions from 1) organization, 2) development, 3) grammar to 4) plagiarism.

Hypothesis 4 in this study is supported by the results from Table 4.13 along with raters' consensus on the foci of their sentence annotating/commenting reported in Table 4.14. It suggests that raters not only have an agreement on score assignment, but also share a common scoring focus when evaluating writing qualities.

CHAPTER 5

DISCUSSION

5.1. Revisit Rater Reliability via Raters' Reading Behaviors

Moss (1994) argued that conventional operationalization of reliability, including rater reliability and task or score reliability, unnecessarily privileged standardized assessment practices over performance based assessment. Therefore, she called for the consideration of a hermeneutic approach, which is a "holistic and integrative approach to interpretation of human phenomena that seeks to understand the whole in light of its parts, repeatedly testing interpretations against the available evidence until each of the parts can be accounted for in a coherent interpretation of the whole" (p.7). This study attempted to explore the potential of a hermeneutic approach proposed by Moss. Instead of focusing on final scores assigned by rater, this study explored the rating process and make interpretations and draw inferences of writing tasks based on raters' scoring behaviors.

Considering the fact that essay raters are text readers at the same time, their scoring decision is naturally affected by their reading behaviors. As raters are presumed to understand the content of the compositions in order to evaluate writing quality, the current research method provides an alternative means to quantify the reliability of raters' scoring decision making and the related impact on test reliability and validity by investigating raters' text reading patterns. The present study examines raters' reading behaviors from several different angles, including reading speed, reading digression-regression rate and attention distribution. The Integrated Rating Environment offers a way to measure such behaviors directly. By doing so, the author is able to study directly the nature of rater reliability as a psychological/behavioral process instead

of building our knowledge about rater reliability on the final scoring result.

The results from the current study indicate that rater reading speed and their reading digression/regression rate can be considered as robust indicators of text comprehension and scoring focus. A fast reading rate and a low digression rate suggest a lack of engagement during reading and hence implying low rater reliability. Rater 1, for example, read the essay at an exceptionally high speed without frequent reading comprehension check. His reading pattern demonstrates a strong potential of lack of attention during essay grading, which explains why rater 1 is associated with a comparatively low inter-rater reliability. On the contrary, if a rater has a high reading regression/digression rate and a relatively low reading rate, it is probable that this rater understands very well the essay content and has a thorough understanding of the writing quality of the text. His reading pattern, in this case, may suggest a higher rater reliability as he would be able to evaluate a composition more precisely and consistently based on the prescribed scoring rubrics. The inter-rater reliability estimated from the scores assigned by the current raters indeed points to the same direction.

Despite the importance of raters' role as text reader in a writing test, their major reading purpose is beyond basic text comprehension. The ultimate goal of their reading is to capture a full range of writing quality of the essays and evaluate the writing based on the scoring benchmarks. There is no surprise that raters should pay more attention to the essay features that are directly associated with the required scoring dimensions. Therefore, when reading the text, raters' reading speed is presumed to fluctuate as they are expected to spend more time processing certain text strings, such as topic sentences, thesis statement and transitional phrases, and scan/skim some essay chunks that are not directly associated with a particular scoring criterion.

This assumption is supported by the results shown in Table 4.3 and 4.5. In this study, the normalized length of all essays was regressed onto the normalized reading time and Table 4.3 provides summary statistics of raters' regression R-square and related reading rates. The larger the R-square is, the larger probability that, however the reading rate is, this rater reads an essay at a constant speed. That is to say, a unit change of his reading time is associated with a unit change of the total essay length. On the other hand, a smaller regression R-square suggests a larger reading digression rate, indicating a larger probability that the rater frequently regress to previous essay chunks or shift his attention to the following or more distant strings. This reading pattern may result in a more fluctuating reading speed; however, it does not necessarily imply a slow reading rate, as we may observe on rater 4 in Table 4.3. Compared to reading rate, the regression R-square as the estimate of raters' reading digression rate is a more robust indicator of rater reliability. The results in Table 4.5 suggest that, regardless of raters' reading speed, a more reliable rater in general demonstrates a larger reading digression rate. This result suggests that reliable raters are able to strategically process a text by capturing the target features prescribed in the rating rubrics. The less reliable raters, however, tend to assign a score based on their truly "holistic" impression of a text, which may vary subjectively.

In this study, raters' reading time is also used to estimate their reading/scoring attention within and across essays. The current results thus provide robust information of the normality of raters' text processing and essay scoring. In this study, the rating normality was based on raters' reading patterns and their scoring behaviors. The "normal" rating process requires a rater to follow a certain reading pattern (relatively low reading rate and high reading digression rate) and have a scoring and reading focus shared by most other raters. Raters' attention distribution was estimated via their reading time spent on particular linguistic units in an essay or certain essay

chunks. In the current investigation, raters' total reading time for each essay is positively correlated with essay features including *total number of words, total number of sub-sentences* and is negatively correlated with number and type of *transitional words*. If this correlation is assumed normal for all raters as a group, the further examination of each rater's reading time for a particular essay would show if an individual rater demonstrates the same reading normality. Along with the correlation between raters' reading time and essay features, raters' scoring foci on certain essay strings or certain scoring dimensions were also estimated via reading time. For example, according to Figure 4.4, most raters spent more time reading the introduction, conclusion and the very middle part of essay 11. If we look into raters' scoring attention across essays, it is evident that their reading time is affected by certain writing qualities of a composition such as organization, content and logical coherence. In this case, reading time will be a robust indicator of readers' attention distribution as we observed from Figure 4.4.

Besides the rough distribution of scoring attention on different parts of an essay, this study provides a text-based attention display to visualize raters' attention distribution within an essay. By visualizing the attention "hot spot" (defined as sentences/phrases that attract more reading time) on each essay, we are able to directly examine the text chunks that readers paid attention to and further analyse features of the "hot spot". The current results show that the distribution of raters' attention "hotspot" (hence, raters' scoring foci) can be categorized into 1) thesis statement and adjacent chunks; 2) topic sentence; and 3) sentences carrying transitional devices. These findings can be considered as the reading "normality" indicators, which provide a quality control tool to examine rater reliability. The fact that most raters focus on certain essay features and writing qualities implies the existence of behavioural agreement and consistency when raters make their scoring decisions. If a rater does not pay attention to those features that

are expected to the shared scoring foci, the reliability of this rater may be jeopardized. In this way, beyond statistical analysis based on raters' scoring judgement, rater reliability can be studied directly by capturing the shared scoring foci among raters and hence directly looking into rater agreement/consistency on his text reading and scoring decision making. A comprehensive analysis of raters' reading patterns and their scoring attention/focus distribution at text base would further provide a more thorough interpretation of rater disagreement; with regard to both their final score assignments and their scoring decision making process.

5.2. Raters' Decision Making: Online Data versus Self-Reported Data

Besides raters' reading time, reading digression rater and attention distribution, another two factors were used to examine their scoring behaviours in the holistic scoring of EPT: raters' verbatim annotation and their scoring comments. In this study raters' annotation and comments were categorized into either positive or negative scoring evidences. Results suggest that the ratios of positive/negative annotations and comments for each essay are significantly correlated with the average score assigned by all raters. In other words, a rater tends to leave more negative comments and annotations to an essay associated with a low score. This result suggests that raters' decision making is reflected not only in their score assignment, but also in their scoring behaviours such as annotating and commenting.

Rating comments were categorized into five scoring aspects including 1) essay organization, 2) essay development, 3) grammar and lexical choice, 4) plagiarism and 5) extra-rubric qualities. This study assumed that the amount of commentary/annotations can be viewed as a measure of perceived importance of a certain scoring dimension. A further investigation of the content of raters' annotation and comments demonstrates that raters pay more attention to

essay features that are associated with certain scoring dimensions. According to Table 4.6, raters' comments were most closely related to the scoring criterion of *essay development*, followed by *grammar*, *plagiarism*, *extra-rubric qualities* and *essay organization* as the most important to the least important. The verbatim annotations were also classified roughly into the five categories and the same focus on *essay development* was also identified in the analysis of raters' verbatim annotation. The number of comments associated with grammatical/lexical errors is ranked as the second largest, indicating that grammar and lexis were also viewed as a fundamental scoring criterion to determine an essay score.

This result, however, is quite contradictory to either the instructions that raters received in the pre-scoring training session or their self-reported scoring focus in the post-rating questionnaire. For example, rater 1, 6 and 8 reported that the role of *grammar and lexical choice* should not be overemphasized during essay grading as they ranked it as the least important scoring dimension (see Table 4.14). Their scoring comments, however, demonstrate a strong tendency that these raters searched for grammatical errors when reading essays as they left a large amount of comments addressing grammar errors. In the training/recalibration session, however, raters were instructed to attend to the scoring dimensions of *text organization* and *essay development.* This instruction was designed based on raters' teaching and EPT grading at UIUC, where they taught ESL academic writing courses to international students. In their writing classes, English writing is taught for academic purpose (EAP) rather than English for specific purposes (ESP). That is to say, the writing tasks students have are highly contextualized within an academic setting. The major purpose of these classes is hence to teach student the writing skills that qualify them as a researcher or scholar in their own field of study. As teaching

grammar and lexis is not the primary objective in these courses, teachers are not expected to focus on the correction of formal errors when evaluating students' writing assignments.

There are three possible interpretations of raters' excessive interest in grammatical and lexical features. The first interpretation is that it may just be that grammar and lexical features necessitate more and longer commentary. It might be easier for a rater to explain his perception of grammar and lexis than to explain perception of other global features such as the organization and idea development of an essay. This conclusion, however, is not supported by previous studies in teacher/rater commentary in either L1 or L2 context. Studies on teacher commentary on English composition reported that writing evaluative commentary is one of the great tasks composition teachers share, and hence it has been one of the central areas of examination in composition studies. However, when L1 and L2 composition raters are asked to articulate their scoring criteria via scoring comments, inconsistency and unevenness in evaluation become apparent across raters (Brown, 1991; Kobayashi, 1992; Leki, 1995; Prior, 1995). As Devenney (1989) pointed out, according to raters' scoring commentary, no group of raters can be completely homogeneous in terms of the qualities they value in students' writing. While some raters focus principally on substance, rhetorical structure, and writing style, others regularly aim at mechanical concerns such as sentence grammar, spelling, and punctuation (Gungle & Taylor, 1989). The fact is that most raters probably invoke a unique combination of these criteria and assign different priorities to a number of these concerns.

Connors and Lunsford (1993) conducted a large scale analysis of teacher commentaries on students' compositions. Their major research objective was to study the patterns and features of comments that address either formal errors or global comments in response to the content of the paper or to the specifically rhetorical aspects of its organization. This study found that raters

showed a balanced attention in their scoring commentary to both global and formal features in the compositions that they assessed. The results of their finding are reported in Table 5.1.

Table 5.1: Numerical Results: Global Commentary Research (Connors & Lunsford, 1993).

Table
Numerical Results: Global Commentary Research

Total number of papers examined: 3,000	# of 3,000	Percentage
Number of papers with global or rhetorical comments	2,297	77% of all Ps
Papers without global or rhetorical comments	703	23%
Number of papers graded	2,241	75%
Number of papers with initial or terminal comments	1,934	64%
Number of initial comments	318	16% of Ps with I or T comments
Number of terminal comments	1,616	84% of Ps with I or T comments
Purpose of comments:		
To give feedback on draft in process	242	11% of Ps with I or T comments
To justify grades	1,355	59%
Global comments in general		
Comments that are all essentially positive	172	9% of Ps with I or T comments
Comments that are all essentially negative	451	23%
Comments that begin positively and then go to negative	808	42%
Comments that begin negatively and then go to positive	217	11%
Comments that lead with rhetorical issues	692	36%
Comments that lead with mechanical issues	357	18%
Very short comments—fewer than 10 words	460	24%
Very long comments—more than 100 words	101	5%
Comments focused exclusively on rhetorical issues	472	24%
Comments focused exclusively on formal/mechanical issues	435	22%
Comments that argue with content points made in paper	478	24%
Comments that indicate use of mechanical criteria as gate criteria ("The comma splices force me to give this an F despite. . . . ")	150	8%
Comments that give general reader response ("like/dislike")	322	17%

Table
(Continued)

Total number of papers examined: 3,000	# of 3,000	Percentage
Comments evaluating specific rhetorical elements:		
Supporting evidence, examples, details	1,296	56% of all Ps with comments
Organization	643	28%
Purpose	240	11%
Response to assignment	246	11%
Audience	137	6%
Overall progress, beyond commentary on paper	176	8%
Comments that deal with specific formal elements:		
Sentence structure	767	33% of all Ps with comments
Paragraph structure	417	18%
Documentation	154	7%
Quotations	142	6%
Source materials	133	6%
Paper format	372	16%

Among 3000 experimental papers, they found that 77% contained global comments. Around 24% comments focused exclusively on rhetorical issues and 22% on formal/mechanical issues. The categorization of specific essay elements in Connors & Lunsford's study was not 100% aligned with the categorization in their investigation. Among the formal elements, it was "sentence structure" that partially represents the "grammar and lexis" scoring dimension in the present scoring rubrics. As the most widely noted formal feature, this element was mentioned in 33% of the commented papers. Since "sentence structure" did not merely refer to syntactic or grammatical complaints or corrections but longer comments on the effectiveness of sentences, the actual comments on pure syntax or lexis should occur in less than one third of all commented papers. The categories of "supporting evidence, examples, details" in Table 5.1 is a subset of the scoring dimension of "essay development" in the present scoring rubrics. A full 56% of all papers with global comments contained comments on the effectiveness or the lack-of supporting

details, evidence, or examples. The next most commonly discussed rhetorical element, at 28%, was overall paper organization, especially issues of introductory sections and issues of conclusion and ending, and thematic coherence.

These results in Table 5.1 surprisingly coincide with findings in the present study. The rank order of number of comments addressing “supporting evidence, examples, details” and “organization” is identical to that of two scoring criteria “essay development” and “essay organization”. The lengths of comments show a large variation. The longest comment they found was over 250 words long, but long comments were far less common than short. Very short comments fewer than ten words were much more common than longer comments. A full 24% of all global comments had ten words or fewer; of these, many were a very few words, or one word- such as "Organization" or "No thesis". There is no strong evidence that grammar and lexical features in the essays generate more and longer commentary. Based on their results as shown in Table 5.1, it is also plausible to conclude that raters tend to address both formal and global issue when leaving essay commentaries, and more global comments are more frequently associated with essay features about text organization and idea development.

A second interpretation of some rates' focus on grammar and lexis is that raters' language background and their teaching and learning experience may make their attention attend to certain essay features. For example, non-native speakers of English may be exposed during their English learning experience to a larger and richer field of technical jargon regarding lexis and grammar than regarding idea development. Therefore, those ESL/EFL raters might feel more comfortable to leave commentaries associated with form-based errors. This hypothesis is partially supported by previous studies of essay raters' decision making process. Cumming et al (2001) documented three coordinated exploratory studies that developed empirically a framework to describe the

decision making of experienced writing raters when evaluating ESL/EFL compositions. They found raters pay more attention to rhetoric and ideas in compositions they scored high than in compositions they scored low, as appose to language features. The ESL/EFL raters attended more extensively, though, to language than to rhetoric and overall ideas, whereas the English-native-speaking (ENS) raters balanced more evenly their attention to these features of the written compositions.

Results from the current study, however, suggest different conclusions. Both ESL/EFL raters and ENS raters have demonstrated unexpected interest in grammatical and lexical features in essay commentaries. Among the five raters who left most language-related comments, three of them are EFL raters and two are ENS raters. The current results show no significant difference between the amount of language or idea comments left by ESL/EFL raters and ENS raters. Therefore, in the present study, it is plausible to conclude that raters' native language background is not a primary factor that influences raters' scoring commentary focus. If we compare the comments left for essays scored high and low, we can find that raters tend to leave more negative comments in essays with a low score than essays with a high score. The current results also suggest that raters left a larger amount of commentary addressing ideas when grading essays that was given a high score. The different commentary foci among raters were also observed, yet this disagreement occurred between experienced and inexperienced raters rather than between ESL/EFL and ENS raters.

It seems that raters' extensive focus on grammar and lexis in an essay could not be accounted for by raters' language background or their teaching experience, or by the nature of grammar and lexis that necessitate more and longer commentary. The current work proposes a third interpretation: the large amount of commentaries on grammar and other language features

may be accounted for by raters' training and scoring experiences. In this study, the number of grammatical and lexical comments was not evenly balanced among raters. Only a certain group of raters that were extensively interested in this scoring dimension during essay commenting. In the rater-recalibration session before the current data collection, all raters were instructed to focus on global features in a text such as organization and essay development. Nevertheless, five raters, rater 5, 6, 8, 9 and 11, still left a large number of comments that are closely related to grammatical features. All of these raters were relative inexperienced ESL TAs who had not been trained to grade EPT essays before the experiment. Therefore, these raters' unusual attention to grammatical and lexical features in an EPT essay could be explained by less training experience and their lack of operational EPT grading experience.

Last but not least, the fact that the discrepancy occured between raters' online scoring behaviours and their self-reported information implies that raters' self-reported scoring focus/attention may not be consistent with their actual scoring behaviours. In other words, raters' retrospective report on how they arrive at their scoring decision may not be an accurate reflection of their decision making process. Due to the fact that what raters believe they do is not necessarily what they actually do, the current research methodology may provide supplementary information to survey studies or studies adopting think-aloud method that are based exclusively on rater's subjective opinion and hence open a new window for studies of test validity.

Raters' moment to moment scoring behaviors also provide useful information for the design or modification of scoring rubrics. Cumming et al (2001) conducted a comprehensive study of raters' decision making by collecting raters' responses in survey questionnaires or raters' think-aloud protocols. They found that raters focus on certain essay qualities when grading an English composition. When asked what three qualities they believed make for especially

effective writing in the context of a composition examination, the raters responded with various related terms. The text qualities that they most frequently mentioned were: (1) rhetorical organization; (2) expression of ideas, including logic, argumentation, clarity, uniqueness, and supporting points; (3) accuracy and fluency of English grammar and vocabulary; and (4) the amount of written text produced. That the participants were able to identify and distinguish these criteria with some uniformity may suggest that these criteria are of fundamental importance and are concepts both conventional and common to ESL/EFL assessment practices. The definitions of the first two text qualities in their study are similar to the scoring dimensions of "organization" and "idea development" in the present study. The fact that both these essay qualities received more attention among raters implies that these two scoring dimensions should be incorporated in the designing of scoring rubrics for an ESL academic writing assessment (TOEFL test in the study of Cumming et al and EPT in the present study). The other two essay qualities, "grammar and lexis" and "essay length" were less frequently mentioned by essay raters according to their answers to survey questionnaires. As this study has suggested an inconsistence between raters' self-reported scoring focus/attention and their actual scoring behaviours, it is necessary to apply the current research methodology to a more comprehensive study targeting at the essay qualities that raters focus on during essay grading. The analysis of raters' natural scoring foci based on their on-line scoring behaviours may provide insights or evidences to the validation of scoring rubrics.

To sum up, a major advantage of this study is to propose indicators beyond test scores that are able to tap directly into raters' decision making process and hence provide alternative methods to estimate the reliability and validity of a writing test. Compared to other indicators of raters' decision making (final scores or think-aloud transcripts), these new indices (e.g. raters'

reading digression rate, reading speed and the ratio of positives/negative comments or annotations) are estimated from the online data collected from raters' decision making process, thus they represent a more accurate reflection of how raters arrive at their scoring decision. The think aloud method is also a good attempt to capture the online record of raters' decision making. However, this method may generate an artificial scoring process as speaking-during-grading is not a natural part of rating process and the think-aloud behavior may even interfere with rater's decision making. Compared with the tedious manual transcription of the think aloud data, data processing in this study is faster and easier as it is automated.

5.3. Integrated Rating Environment: Advantages of the Current Research Instrument

In reading studies, eye trackers have been used to capture features of readers' eye movement, including gaze durations, saccade lengths, and occurrence of repressions, to draw inferences of moment-by-moment cognitive processing of a text (Just & Carpenter, 1980). Compared to traditional studies that ask participants to read on paper, the eye tracking methodology doesn't interrupt the natural reading process and provides moment-to-moment eye movement data with great speed and precision. Therefore, it has been used as an important source of language processing in reading studies. However, eye tracking as a data collection method has its own limitations.

First of all, this method is more costly as compared to other data collection methods. The researchers who use eye tracking technology must be trained on how to use the equipment and may need technical support to help participants set up and get calibrated with the device during data collection.

In addition, eye tracking doesn't provide information about the success or failure of

comprehending a text. Thus, the eye-tracking data must be complemented with other performance measures, such as retrospective comprehension tests or cognitive interviews, which will increase the data collection burden for participants.

Thirdly, it is difficult to code and analyze eye tracking data, which may require the use of specific software. To interpret eye tracking data, the researchers much choose from a list of dependent variables or metrics to analyze in the data stream and these metrics, such as fixation duration and gaze duration, are not quite self-explanatory. Assumptions and inferences must be made when analyzing the eye tracking data and again these data need to be supplemented by other performance measures.

In the current study, the Integrated Rating Environment (IRE), a Python-based rating interface, was used was used as the primary tool to deliver the written samples to the raters and collect their moment to moment scoring data and their post-rating survey answers. The IRE has many advantages compared with other methods of data delivery and data collection.

First of all, the current Rating Environment allows raters to not only assign a score to an essay, but also select and annotate phrases/sentences from the sample writing during their decision-making process. This function helps language testers to explore raters decision making by looking at the online data instead of the final score assignment. While other methods such as think aloud method have also made the effort to collect online rating data, the IRE minimizes the interference to the naturalness of grading process. The extra effort for raters to comment, annotate and assign scores in IRE during essay grading is relatively small after short training and hence has a relatively small impact on their rating decision making. The 'select-highlight' method used to collect reading pattern is not the most natural way for text reading, however most raters seem comfortable to this feature after a short introductory period. While the “observer's

paradox” can never be completely resolved, the current research instrument performs better than most other current research instruments.

Secondly, the IRE makes the scoring collection and analysis automatic. All the events are recorded into a log, which can be used as a source to automatically extract scoring data and annotation data. As part of the IRE, the analysis components make the data extraction automatic. No tedious transcription of oral speech or hand-writing is needed and thousands of scoring events are extracted and organized precisely within milliseconds. This rating interface also enables researchers to visualize patterns or distributions of raters' dynamic online scoring behaviors, such as their reading pattern and attention distribution over the texts.

Finally, it is also more cost effective for long distance data transfer and data delivery. The rating interface with the essays to be rated can be uploaded to and downloaded from a website. Therefore, the IRE saves shipping time and expenses. In addition, the automatic data extraction in the rating interface also avoids possible coding errors in the traditional method of essay grading and data collection.

Though the IRE was designed for the study of ESL rating, this rating interface can be applied in different writing contexts; therefore, the indicators generated in the present study are not limited in the EPT writing test. The current study can be then expanded to examine essay raters' decision making process in other writing assessments that are of different test scales, different rating rubrics and different scoring dimensions, for example, IELTS or TOEFL.

CHAPTER 6

CONCLUSIONS

6.1. Findings and Limitations of the Current Study

In the current study, the ITM framework was adopted to investigate raters' decision making process for the EPT writing test at UIUC. This study looks into the construct validity of the new version EPT from the perspective of raters' decision making process. The purpose of this paper is thus to evaluate if the Semi-Enhanced EPT measures the target construct and if raters' scoring behavior is consistent in their own grading or across different raters. This study also serves to test four research hypothesis noted below.

Hypothesis 1: *A high reading digression rate and a low reading rate indicate an engaged reading comprehension process during essay grading, hence these indices are positively associated with rater reliability in a writing test.*

Hypothesis 2*: If there is an interaction between rater and essay writer, raters' scoring decision is associated with essay features.*

Hypothesis 3*: Raters' decision making is reflected not only in their score assignment, but also in their scoring behaviours such as sentence selection, verbatim annotation and comment.*

Hypothesis 4*: Raters not only have an agreement on score assignment, but also share a common scoring focus when evaluating writing qualities.*

The current research findings support all these four hypotheses. In this study, raters had a common scoring attention (calculated from their text reading time), which is distributed according to essay features related to prescribed scoring criteria (e.g. *essay development*). Raters

also shared a common focus on the development criterion during essay commenting. Their positive comment hotspots clustered around thesis statement, topic sentences and transitional devices. On the other hand, the negative hotspots are more of content level, putting more emphasis in essay development than other scoring criteria. These findings partially support that the SEEPT raters in fact evaluate the students' academic writing ability based on required scoring dimensions, thus enforcing the construct validity of the test.

A strong rater-essay interaction has been observed in this study, indicating that raters' scoring decision making is affected by their text reading and also essay features. Raters' reading time is correlated with various essay features: it is positively correlated with number of vocabulary, essay length, the number of sentence and subsentence; and negatively correlated with the number and category of transitional devices. Most raters demonstrate a linear reading pattern during their text reading and essay grading. A rater-text interaction is further supported by the correlation between essay scores and text features: essay score is positively correlated with # of vocabulary, sentence length and transitional devices. Essay score may be negatively correlated with word frequency.

Raters' self-reported data is not consistent with their scoring behaviors. Their sentence annotation and scoring comments demonstrate different scoring focus comparing to their answers to post-grading survey questions. This finding demonstrates a limitation in previous research methodologies -- raters don't behave as they said or as they thought they would. A difference between trained rater and untrained rater is also identified in this work. Compared to experienced raters, untrained raters tend to over emphasis the importance of "grammar and lexical choice".

Another purpose of the current study is to develop empirically an exploratory framework that describes essay raters' decision-making processes while holistically rating compositions in

an integrated writing performance test, e. g. the EPT writing test. Findings from the current investigation implies that this purpose has been achieved via the descriptive analysis of raters' reading patterns, their reading attention and raters' scoring focus on certain essay qualities. As the status of this research remains exploratory, further studies with more rigorous empirical means, different populations, writing tasks, conditions for writing, and methods of inquiry would help to verify and refine the proposed framework. With such future work, the present descriptive framework may serve as a fundamental pre-cursor to future new models that specify or evaluate procedures for rating ESL/EFL writing performance tasks in different test contexts.

Generally speaking, raters' reading and scoring behaviors represent their scoring process and interrelated decisions that composition raters are expected to make routinely while they holistically rate essay samples in ESL/EFL writing assessments. These behaviors are worth considering as benchmarks of decision making in designing schemes for scoring ESL writing; providing instructions to guide raters; selecting, rating, or monitoring raters; creating checklists of desirable behaviors for raters to use or learn to develop; identifying behaviors that might not be desirable for specific assessment purposes; or conducting future research on this topic. Moreover, findings from this research indicate specific aspects of decision making where standardization or training of raters may be able to improve raters' reliability or consistency while scoring ESL/EFL composition.

Like previous research on raters' decision making processes, the present study find that the evaluation of ESL/EFL compositions involve interactive multifaceted decision making. Fundamentally, the raters balance processes of interpretation with processes of judgment while attending to numerous aspects of essay qualities. These cognitive processes operate in conjunction with criteria or values that experienced raters necessarily use to guide their holistic

scoring of writing samples. The rating tasks for the present research specify the scoring criteria in advance and raters also share a similar teaching and grading experience in the same ESL program, so the raters have to rely on both their accumulated knowledge from prior experiences in assessing essays and their familiarity to the scoring benchmarks to guide themselves in attributing scores to the writing samples. During the essay grading, each rater was given the scoring benchmarks and the recalibration essays so that they were able to check the expected performance for each scale (placement) level. Most experienced raters, however, only referred to these recalibration materials once or twice, indicating that while they rated the compositions they have established the internalization of specific scoring criteria or they were able to recall criteria or benchmark situations from their previous EPT grading experience. These findings may usefully reflect prevailing educational norms as well as the accumulated, relevant experiences that experienced raters possess. Therefore, the holistic schemes for rating ESL compositions may necessarily require precise criteria as to the levels of performance expected of examinees on particular tests and tasks in order to assure validity in the specific testing environment.

This research also makes suggestions in designing and modifying scoring criteria for assessing ESL/EFL writing performance. The experienced raters participating in the present study all showed a proportional balance in their decision making between attention to rhetoric and ideas and to language features in the ESL/EFL compositions that they assessed. This finding implies that when grading essays holistically, raters still assess writing qualifies by evaluating specific essay features in multiple scoring dimensions. Indeed, analytic scales corresponding to each of these scoring dimensions may more realistically represent how experienced raters conceptualize ESL/EFL writing proficiency than, for example, a single holistic scale that combines these dimensions as in the current scale for the EPT essay. Due to the placement

purpose of the EPT writing test, analytic scales may also provide useful diagnostic information for the ESL instructors.

Results from the current study also suggest reasons to weigh criteria differently toward certain essay aspects at different placement levels of a rating scale. It seems that raters' grammar and lexis related comments are primarily associated with lower-scored essays. The essays at the higher end, however, obtained more comments associated with rhetorics and ideas. This finding implies that language aspects needs to be more heavily weighted at the lower end of a rating scales, while global features should be focused at the higher end. The fact that most raters attended more to language than to global features on essays they graded low indicates that adult ESL/EFL learners may have to attain a certain threshold level in their language abilities before raters can attend thoroughly to the ideas and rhetorical abilities in compositions.

The overall behavioral evidence for raters' decision making suggests that experienced ESL raters' decision making might be fundamentally similar across different types of writing tasks, however, they probably still need unique criteria for scoring particular types of writing with a particular purpose. Indeed, most experienced raters in this research were so familiar with the scoring benchmarks due to their previous EPT grading experience, but some less experienced raters found that they needed explicit guidelines to know how to evaluate examinees' performance even though they have graded compositions of their ESL students by using a very much similar scoring benchmarks. However, in their own ESL academic classes, they grade composition to assess students' English writing proficiency while in the EPT writing test, these inexperienced EPT raters are supposed to evaluate students' writing qualities for placement purpose. These different scoring purposes determine that the raters who did not have operational scoring experience may demonstrate different scoring foci as we observed in this study.

In a related way, this study has confirmed that groups of raters with common professional or educational backgrounds act in reference to certain norms and expectations, as has been shown in previous inquiry comparing the behaviors of differing groups of raters of ESL compositions. However, differences in decision-making processes across groups of raters may not be as great as such other studies have founds when analyzing their ratings of essays alone. For instance, the ESL/EFL raters and ENS raters displayed fundamentally the same decision making behaviors when rating comparable EPT essays. However, this conclusion probably only makes sense within the limited discourse community of a particular program at a specific educational setting, rather than in reference to the great diversity of different text contexts.

Limitations of the descriptive framework also need to be considered. The fundamental question that hasn't been answered in this study is to what extent decision making behaviors can be generalized and standardized to evaluate if a rater's scoring is reliable. Due to a small convenience sample and the descriptive nature of the study, results from the current work cannot be generalized to a larger population of essay graders. Therefore, it would be premature to conclude that the common behavioral patterns shared by experienced raters may provide precise benchmarks to evaluate if a rater is reliable or not. It would be more appropriate to use the current results as quality control tools for rater monitoring and rater training. By comparing raters' reading and scoring behaviors to the shared group behaviors, we may identify those raters at risk and then take further actions before an unreliable rater jeopardizes the validity of this writing test. Additional statistical analysis, such as a generalizability study, may also provide useful information to test developers in terms of test dependability and possible source of measurement error.

In addition, the descriptive indicator of raters' decision making, e.g. reading time, reading

digression rate and ratios of positive/negative annotations and comments, have their own limitation. As these factors are newly applied in the study of raters' decision making in the current study, a further validation of these indicators may be necessary in a study of larger scale. At current stage, there is not existing formula or statistical package which can be used to test the significance of the normality of these indicators across different test contexts. In other words, there is no fixed standard or cut-off value for the result interpretation of these indicators and these factors are all case sensitive. More works are needed to validate the estimation of these indices and further investigate the sense-of-baseline. Due to the limited amount of data collected in the present study, the employment of these indicators of raters' decision making in large scale studies across different scoring dimensions is subject to necessary validation of the effectiveness of these indicators in writing assessment.

Last but the not the least, the utility, clarity, and accessibility of the IRE should be further evaluated and refined. For example, the current interface doesn't document the comments deleted by users or any changes of assigned essay grades. Feedback from users of the interface and computer interface developers should be collected to review the current functions of the IRE and make further modifications. In order to capture the full spectrum of graders' essay comprehension and decision making process, eye tracking techniques may also be used in future studies to complement the use of one manual input device.

6.2. Future Studies

Due to the limitation of time frame and resources, many topics regarding the rating process of ESL writing performance assessments are not discussed in this study. However, this study provides the methodological means to the validation of writing performance assessments.

Test validation, referred as a broad spectrum of empirical data collection activities, may yield evidence to justify using test scores for making specific types of inferences about examinees. According to Miller and Crocker (1990), language testers have conducted validation studies to answer the following questions:

1. Does the writing exercise adequately represent the content domain?

2. Do different scoring procedures applied to direct writing assessments yield similar results (i.e., measure the same trait)?

3. Do direct and indirect measures of writing yield similar results (i.e., measure the same trait)?

4. Can writing samples be used to predict external criteria (e.g., course grades)?

5. What extraneous factors may influence examinee performance or ratings assigned to the writing sample?

Each type of these investigations exemplifies a specific type of validation operation in the overall process of construct validation set forth by Messick (1989). According to this schema, language testers in test validation do not examine the validity of test content or test scores themselves, but rather the validity of the way we interpret or use the information gathered through the testing procedure.

In the current research target, the writing assessments, a fundamental question to be answered in validation is that if raters accurately and consistently evaluate compositions based on the prescribed benchmarks. Due to the subjective nature of the scoring process in a performance based wring test, the “rating validity” directly determines if the test is actually evaluating the target writing abilities of the test takers or some other factors introduced in the

rating process. Within the current framework, a new approach is applicable to the investigation of “rating validity” by the micro analysis of raters' decision making behaviors in rater training and their operational scoring.

In writing tests, raters' scoring judgment was typically quantified and evaluated using a rating scale. One of the basic questions that arise in these situations is how to evaluate the quality of subjective judgments obtained from raters. Therefore, rater accuracy and consistency have been a long-term research interest among scholars and test experts. Most studies, however, examine rater accuracy or consistency within statistical frameworks by addressing raters' final score assignment only. For example, Engelhard (1996) defined rater accuracy as the match between the ratings obtained from operational raters and the ratings assigned by an expert panel to a set of benchmark or exemplar performances, therefore, the higher the correspondence between the operational and benchmark ratings, the higher the level of rater accuracy. Within the current research framework, rater accuracy and consistency can be examined by directly investigating the correspondence between the actual scoring behaviors of both operational raters and expert raters. By using the current research instrument, the rating interface, the behavioral patterns of expert raters could be monitored and standardized to evaluate the accuracy and consistency of operational raters. For example, within the context of large-scale ESL writing assessment, e.g. TOEFL *ibt* writing, a set of student papers from the field test or an earlier administration of the assessment can be selected as benchmarks. These benchmark papers can then be rated both by an expert panel and by operational raters, and the match between operational and benchmark ratings can be used as an indicator of rater accuracy. The closer the behavioral correspondence between the operational ratings and the benchmark ratings, the higher the level of accuracy. The rater consistency then can be defined as the level or degree of

behavioral consistency an individual demonstrates comparing to his previous ratings or peer ratings. This new approach of examining rater accuracy and consistency then provides more precise understanding of how and why a rater arrives at a particular scoring decision.

The current study also provides useful feedback to rater screening and rater training, as the results of rater accuracy can be used in rater training programs to screen out inaccurate raters, to provide feedback to inaccurate raters, to monitor the ongoing quality of raters over time, and to evaluate the influences of rater training. In the development of an operational performance assessment system using accuracy indices, there are a variety of substantive issues that need to be addressed in future research. First of all, there are several questions related to the selection of benchmark performances. How should the benchmark performances be selected? Should the benchmarks be uniformly distributed over the scale or not? How should the reliability of the benchmark ratings be determined via raters' scoring behaviors? Next, it is important to consider how to actually use the benchmark performances within an operational assessment system. How accurate do raters have to be in order to be considered accurate enough to begin or to continue rating? Is a "cut-off score" needed to define acceptable rater accuracy? If so, how should this value be determined based on indicators representing raters' scoring behaviors? How stable are the behavior estimates of rater accuracy over time? Will raters' reading and rating behaviors change over time or across different writing prompts? Last but not least, future research is also needed on the amount and kind of feedbacks that should be provided to operational raters based on the evaluation of their rating accuracy and consistency.

REFERENCES

Adams, R.J., Griffin, P.E. & Martin, L. (1987). A latent trait method for measuring a dimension in second language proficiency. *Language Testing* 4(1), 9–28.

Altarriba, J., Kroll, J., Sholl, A. & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory & Cognition*, 24, 477-492.

Anderson, N., Bachman, L.F., Cohen, A.D. & Perkins, K. (1991). An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources. *Language Testing* 8(1), 41–66.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F., Lynch, B.K., and Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*, 238-257.

Bachman, L.F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L.F. & Eignor, D.R. (1997). Recent advances in quantitative test analysis. In Clapham, C. and Corson, D. (Eds.), 227-242.

Bachman, L.F. (1998) Appendix: Language testing –SLA interfaces. In L.F. Bachman, & A.D. Cohen, (Eds.), *Interfaces between second language acquisition and language testing research*, pp 1-27. Cambridge: CUP

Bachman, L.F. (2000). Modern language testing at the turn of the century: Assuring that what we count. counts. *Language Testing, 17*, 1–42.

Bailey, B. (1999) *UI Design Update Newsletter , February, 1999.* [On-Line] Available: http://www.humanfactors.com/ library/feb99.asp

Bamberg, B. (1983). What makes a text coherent? *College Composition and Communication, 34, 417-429.*

Bauer, B. A. (1981). *A study of the reliabilities and cost efficiencies of three methods of assessments for writing ability.* Champaign: University of Illinois (ERIC Document Reproduction Service No. 216 357).

Berry, V. (1993). Personality characteristics as a potential source of language test bias. In Huhta, A. Sajavaara, K. and Takala, S., (Eds.), *Language testing: new openings*. Jyva¨skyla: University of Jyva¨skyla, 114–124.

Bolus, R.E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning, 32*, 245-258.

Bolt, R.F. (1992). Crossvalidation of item response curve models using TOEFL data. *Language Testing* 9, 79–95.

Braze, D., Shankweiler, D., Ni,W., & Palumbo, L. C. (2002) Readers' eye movements distinguish anomalies of form and content. *Journal of Psycholinguistic Research*, 31: 25–44.

Breland, H., & Jones, R. J. (1984). Perceptions of writing skills. *Written Communication, 1, 101-109.*

Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research (pp. 112-140).* Cambridge: Cambridge University Press.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing* 12(1), 1–15.

Brown, J.D. (1999). Relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16* (2), 216-237.

Brown, J.D. and Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly* 32(4), 653–75.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1(1*), pp.8-24.

Canale, M., (1983). From communicative competence to communicative language pedagogy. In *Language and Communication* J.C.Richards & R.W.Schmidt(Eds.). Longman, Harlow, p.2-27.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research on the Teaching of English, 18*, 65-81.

Chinn, J. A. (1979). Verb choice and its effectiveness as measured by holistic evaluation. (ERIC Document Reproduction Service No. 198 545).

Clapham, C. (1993). The effect of language proficiency and background knowledge on EAP students' reading comprehension. In Kunnan, A.J., (Ed.), *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum, pp. 141–68.

Clapham, C. (1996). *The development of IELTS: a study of the effect of background knowledge on reading comprehension*. Cambridge: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Congdon, P. J. & McQueen, J. (2000). The permanence of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37, 163-178.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? TESOL Quarterly, 29 (4), 762-765

Cook,A.E. (2005).*Failures to detect inconsistencies inanaphoricreferences*.Manuscriptin preparation.

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizbility: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137-163.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. NY: Wiley.

Crowhurst, M. (1980). Syntactic complexity and teachers' ratings of narrations and arguments. *Research in the Teaching of English, 14, 223-231.*

Cumming, A. (1990) Expertise in evaluating second language compositions. *Language Testing*, v7 p31-51.

Cumming, A. (1997). The testing of writing in a second language. In Clapham, C. and Corson, D., editors, *Encyclopedia of language and education*, 7, 51-53. *Language testing and assessment*. Dordrecht, Netherlands: Kluwer.

Cumming, A., Kantor, R., and Powers, D. (2001). Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework (TOEFL Monograph No. 22). Princeton, NJ: ETS.

de Jong, J.H.A.L. (1986). Item selection from pretests in mixed ability groups. In Stansfield, C.W., editor, *Technology and language testing*. Arlington, VA: TESOL, 91–108.

Deutsch, A., & Bentin, S. (2001). Syntactic and semantic factors in processing gender agreement in Hebrew: evidence from ERPs and eye movements. *Journal of Memory and Language*, 45: 200–224.

Diederich, P. B. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability (*Research Bulletin 61-15*). Princeton, NJ: Educational Testing Service.

Du, Y., & Wright, B. D. (1997). Effects of student characteristics in a large-scale direct writing assessment. In M. Wilson, G. Engelhard Jr., & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 1–24). Stamford, CT: Ablex.

Duffy, S., & Keir, J. A. (2004). Violating stereotypes: eye movements and comprehension processes when text conflicts with world knowledge.*Memory and Cognition*, 32: 551–559.

Ehrlich, K., & Rayner, K. (1983). Pronoun assignment and semantic integration during reading: eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, 22: 75–87.

Emig, J. A., & Parker, R. P., Jr. (1976). *Responding to student writing: Building a theory of the evaluating process.* Report prepared at Rutgers University. (ERIC ED 136 257).

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31,* 93–112.

Ericsson, K. A. & Simon, H. A. (1993). Protocol analysis: Verbal reports as data. Revised edition. Cambridge, MA: The MIT Press.

Fahenstock, J. (1983). Semantic and lexical coherence. *College Composition and Communication, 34, 400-415.*

Faigley, Lester, Roger D. Cherry, David A. Jolliffe, and Anna M. Skinner (1985). *Assessing Writers' Knowledge and Processes of Composing.* Norwood, NJ: Ablex.

Fortus, R., Corriat, R. & Fund, S. 1(998). Prediction of item difficulty in the English subtest of Israel's inter-university Psychometric Entrance Test. In Kunnan, A.J., (Ed.), *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum, 61–87.

Freedman, S. W. (1979a). How characteristics of student essays influence teachers' evaluation. *Journal of Educational Psychology, 71, 328-338.*

Freedman, S. W. (1979b). Why do teachers five the grades they do? *College Composition and Communication, 30, 161-164.*

Freedman, S. W. (1981). Influences of evaluation of expository essays: Beyond the text. *Research in the Teaching of English, 15, 245-255*.

Freedman, S. W. (1984). The registers of student and professional expository writing. Influences on teachers' responses. In R. Beach & S. Bridwell (Eds.). *New directions in composition research* (pp. 334-347). New York: Guilford Press.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, S. A. Walmsley (Eds.), *Research on Writing: Principles and methods (pp. 75-98).* New York: Longman.

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing* 13(1), 23–51.

Garrod, S., Freudenthal, S., & Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language*, 33: 39–68.

Geva, E. (1983). Facilitating reading comprehension through flowcharting. *Reading Research Quarterly, 17, 384-341*.

Geva, E. (1992). The role of conjunctions in L2 text comprehension**.** *TESOL Quarterly*, Vol. 26, No. 4. pp. 731-747

Gleser, G.C., Cronbach, L.J. & Rajaratnam, N. (1965). Generalizability of Scores Influenced by Multiple Sources of Variance. *Psychometrika*, 30, 345-418.

Golding, J. M., Millis, K. M., Hauselt, J., & Sego, S. A. (1995). The effect of connectives and causal relatedness on text comprehension. In R. F. Lorch, Jr. & E. J. O'Brien (Eds.), *Sources of Coherence in Reading* (pp. 127-143). Hillsdale, NJ: Lawrence Erlbaum.

Goodman, K. S. (1971). Psycholinguistic universals in the reading process. In P. Pimsleur & T. Quinn (Eds.), *The psychology of second language reading* (pp. 135-142). Cambridge: Cambridge University Press.

Goodman, K. S. (1996). On reading. Portsmouth, NH: Heinemann.

Graesser, A. C., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension, *Psychological Review* (101), No. 3, 371-395.

Greenberg, K. (1981). *The effects of variations in essay questions on the writing performance of CUNY freshman*. New York: The City University of New York Instructional Resource Center.

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality rating. *Research in the Teaching of English, 15, 75-85.*

Gyagenda, I. S., & Engelhard, Jr., G. (1998). *Rater, domain, and gender influences on the assessed quality of student writing using weighted and unweighted scoring*. Paper presented at the Annual Meeting of the American Educational Research Association, April 13-17, 1998.

Haberlandt, K. (1982). Reader expectations in text comprehension. In. J.-F. Le Ny & W. Kintsch (Eds.), *Language and Comprehension* (pp. 239-249). Amsterdam: North-Holland.

Haliday, R., & Hasan, D. (1979). *Cohesion in English*. New York: Longman.

Hambleton, R.K. (1989) Principles and selected applications of Item Response Theory. In R.L. Linn (Ed.), *Educational measurement* (pp. 147–200). New York: Macmillian

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff Publishing

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage.

Hamp-Lyons, L., and Kroll, B. (1997). *TOEFL 2000—Writing: Composition, community, and assessment.* Princeton, NJ: Educational Testing Service.

Henderson, J. M, & Ferreira, F. (1990). The effects of foveal difficulty on the perceptual span in reading: Implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 417-429.

Henderson, J. M, & Ferreira, F. (1993). Eye movement control in reading: Fixation measures reflect foveal but not parafoveal processing difficulty. *Canadian Journal of Experimental Psychology: Special Issue on Reading and Language Processing, 47*, 201-221.

Henning, G. (1991). Validating an item bank in a computer-assisted or computer-adaptive test. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 209-222). New York: Newbury House.

Hill, K. (1993). The effect of test-taker characteristics on reactions to and performance on an oral English proficiency test. In Kunnan, A.J., (Ed.), *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum, 209–30.

Holland, N. N. (1968). *The dynamics of literary response.* New York: Oxford University Press.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, 6, 153-160.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item response theory. Homewood, IL: Dow Jones-Irwin.

Hunt, K. (1965). Grammatical structures written at three grade levels. *National Council of Teachers of English, Research Report No. 3*. Champaign, Illinois: National Council of Teachers of English.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60,* 237-263.

Hymes, D. (1972). On communicative competence. In Pride, J. & Holmes, J. (Eds.), *Sociolinguistics.* Harmondsworth: Penguin.

Hyönä, J., & Vainio, S. (2001). Reading morphologically complex clause structures in Finnish. *European Journal of Cognitive Psychology*, 13: 451–474.

Inhoff, A.W., & Rayner, K. (1986) Parafoveal word processing during eye fixations in reading: effects of word frequency. *Perception and Psychophysics*, 40: 431–439.

Jones, B. E. W. (1978). Marking of student writing by high school teachers in Virginia during 1976. *Dissertation Abstracts International*, 38, 3911A.

Just, M. A. & Carpenter, P. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review*, 85: 109–130.

Just, M. A., & Carpenter, P. A. (Eds.) (1987) *The Psychology of reading and language comprehension*. Newton, MA: Allyn and Bacon.

Keenan, J. M., Baillet, S. D., & Brown, P. (1984). The effects of causal cohesion on comprehension and memory. *Journal of Verbal Learning and Verbal Behavior*, *23*, 115-126.

Kieras, D. E. (1985). Thematic processes in the comprehension of technical prose. In B. K. Britton & J. B. Black (Eds.), *Understanding expository text (pp. 89-107).* Hillsdale, NJ: Lawrence Erlbaum.

Kintsch, W. (1974). *The representation of meaning in memory.* New York: John Wiley & Sons.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review,* 95, s163-182.

Kintsch, W. (1994). The Psychology of Discourse Processing. In *Handbook of Psycholinguistics,* M. Gernsbacher (Ed.), San Diego: Academic Press. pp. 721-739.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W. & Van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85 (5),* 363-394.

Kunnan, A.J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. *Language Testing 9,* 3049.

Lee, Y.-W., Gentile, C., & Kantor, R. (2005). *Analytic scoring of TOEFL CBT essays: Scores from humans and e-rater.* Princeton, NJ: ETS.

Linacre, J.M. (1989). *Many-facet Rasch measurement.* Chicago: MESA Press.

Lorch, R. A. F., & Lorch, E. P. (1986). On-line processing of summary and importance signals in reading. *Discourse Processes, 9, 489-496*.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph No. 7*.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley..

Lumley, T. (2000). The process of the assessment of writing performance: The rater's perspective. Unpublished Ph.D dissertation, Department of Linguistics and Applied Linguistics, The University of Melbourne.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing,* 19, 246-276.

Lumley, T. and McNamara, T.F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing* 12(1), 54–71.

Lynch, B.K., Davidson, F. and Henning, G. (1988). Person dimensionality in language test validation. *Language Testing* 5(2), 206–19.

Lynch, B.K. and McNamara, T.F. (1998): Using g-theory and many-facet Rasch measurement in the development of performance assessments. *Language Testing* 15(2), 158–80.

McCulley, G. A. (1985). Writing quality, coherence and cohesion. *Research in the Teaching of English, 19, 269-282.*

McNamara, T.F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing* 8(2), 139–59.

McNamara, T.F. (1995). Modelling performance: opening Pandora's Box. *Applied Linguistics, 16*(2), 159–75.

McNamara, T.F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T.F. (1997). Performance testing. In Clapham, C. and Corson, D., editors, *Encyclopedia of language and education*. Volume 7. Language testing and assessment. Dordrecht: Kluwer Academic, 131–39.

McNamara, T.F. & Lumley, T. (1997).The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140–56.

Messick, S. (1989).Validity. In Linn, R.L., editor, *Educational measurement*. 3rd edn. New York: American Council on Education/Macmillan, 13–103.

Meyer, B. J. F. (1977). The structure of prose: Effects on learning and memory and implications for educational practice. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge (pp. 179-200).* New York: Wiley.

Meyer, B. J. F., Brandt, D. N., & Bluth, G. J. (1981). Use of author's textual schema: Key for ninth graders' comprehension. *Reading Research Quarterly, 15, 72-103*.

Mislevy, R.J. (1993). Foundations of a new test theory. In N. Frederiksen, , R.J. Mislevy,& I. Bejar, (Eds.), *Test theory for a new generation of tests.* Hillsdale, New Jersey:Lawrence: Erlbaum Associates, Publishers.

Morrison, R. E. (1984). Manipulation of stimulus onset delay in reading: Evidence for parallel programming of saccades. *Journal of Experimental Psychology: Human Perception and Performance, 10, 6*67-682.

Myers, J. L., Shinjo, M., & Duffy, S. A. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language, 26,* 453-465.

Neilsen, L., & Piche, G. (1981). The influence of headed nominal complexity and lexical choice on teachers' evaluation of writing. *Research in the Teaching of English, 15, 65-74.*

Neuner, J. L. (1987). Cohesive ties and chains in good and poor freshman essays. *Research in the Teaching of English, 21, 92-105.*

Ni,W., Fodor, J. D., Crain, S., & Shankweiler, D. (1998). Anomaly detection: eye movement patterns. *Journal of Psycholinguistic Research*, 27: 515–539.

Nold, E. W., & Freedman, S. W. (1977). An analysis of reader's responses to essays. *Research in the Teaching of English, 11, 164-174.*

Norris, J.M., Brown, J.D., Hudson, T.D. & Yoshioka, J.K. (1998). *Designing second language performance assessments* (Technical Report #18). Honolulu: University of Hawaii, Second Language Teaching & Curriculum Center.

O'Brien, E. J., Raney, G. E., Albrecht, J. E., & Rayner, K. (1997). Processes involved in the resolution of explicit anaphors. *Discourse Processes*, 23: 1–24.

O'Donnell, C., Griffin, W., & Norris, B. (1967).Syntax of kindergarten and elementary school children. *National Council of Teachers of English, Research Report No. 8.* Champaign, IL: National Council of Teachers of English.

O'Regan, J. K. (1979). Eye guidance in reading: evidence for the linguistic control hypothesis. *Perception and Psychophysics*, 25: 501–509.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, 41: 427–456.

Perkins, K. & Brutten, S.R. (1993). A model of ESL reading comprehension difficulty. In Huhta, A., Sajavaara, K. and Takala, S., editors, *Language testing: new openings*. Jyva¨skyla: University of Jyva¨skyla, 205–18.

Perkins, K., Gupta, L. & Tammana, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language Testing* 12(1), 34–53.

Perkins, K. & Gass, S.M. (1996).An investigation of patterns of discontinuous learning: implications for ESL measurement. *Language Testing, 13*(1), 63–82.

Pollatsek, A. & Rayner, K. (1990). Eye movements and lexical access in reading. In D. A. Balota, G. B. Flores d' Arcais, & K. Rayner (Eds.), *Comprehension processes in reading (pp. 143-164).* Hillsdale, NJ: Erlbaum.

Pollitt, A. & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, (1),* 72-92.

Pollitt, A. (1997). Rasch measurement in latent trait models. In Clapham, C. and Corson, D.,(Eds.), *Encyclopedia of language and education*. Volume 7: Language testing and assessment. Dordrecht: Kluwer Academic,243–54.

Purves, A. C. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English*, 26, 108-122.

Raimes, A. (1990). The TOEFL test of written English: Causes for concern. *TESOL Quarterly, 24*, 427-442.

Raney, G. E. & Rayner, K. (1995). Word frequency effects and eye movements during two readings of a text. *Canadian Journal of Experimental Psychology, 49,* 151-172.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rayner, K. (1977). Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition, 4*, 443-448.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124: 372–422.

Rayner, K., Chace, K. Slattery, T., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading, 10,* 241-256

Rayner, K., Cook, A. E., Juhasz, B. J., and Frazier, L. (2006). Immediate disambiguation of lexically ambiguous words during reading: evidence from eye movements. *British Journal of Psychology*, 97: 467–82.

Rayner, K., & Duffy, S. (1986) Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition*, 14: 191–201.

Rayner, K., & Pollastek, A. (1989). *The psychology of reading.* Englewood Cliffs, NJ: Prentice-Hall.

Rayner, K., Sereno, S. C., and Raney, G. E. (1996) Eye movement control in reading: a comparison of two types of models. *Journal of Experimental Psychology: Human Perception and Performance*, 22: 1188–1200.

Riley, G.L. & Lee, J.F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing* 13(2), 173–90.

Rozeboom, W.W. (1978). Domain validity—Why care? *Educational and Psychological Measurement. 38,* 81-88.

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: quantitative and qualitative analyses*. New York: Peter Lang.

Schoonen, R., Vergeer, M. & Eiting, M. (1997). The assessment of writing ability: expert readers versus lay readers. *Language Testing* 14, 157–84.

Shannon, C. (1951). Prediction and entropy of printed English. *Bell Systems Technical Journal* 30:50—64.

Scherer, D. L. (1985). Measuring the measurements: A study of evaluation of writing. *An annotated bibliography.* (ERIC Document Reproduction Service N0. 260 455).

Shohamy, E. (1983). Rater Reliability of the Oral Interview Speaking Test. *Foreign Language Annals.*16, 3, 219-222.

Shohamy, E. (1984). Does the Testing Method Make a Difference? The Case of Reading Comprehension. *Language Testing, 1*, 2, 147-70.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing* 11(2), 99–123.

Shohamy, E., Gordon, C., and Kramer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, 76 n1 p27-33.

Smith, F. (1971). *Understanding reading: a psycholinguistic analysis of reading and learning to read*. New York: Holt, Rinehart and Winston

Smith, F. (1971). *Understanding reading* (5th edition). Mahweh, NJ: Erlbaum.

Smith, Jr., E. V. & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet rasch measurement using a complex problem solving skills assessment. *Educational and Psychological Measurement, 64*, 617-639.

Sparks, R.L., Artzer, M., Ganschow, L., Siebenhar, D., Plageman, M.& Patton, J. (1998). Differences in native-language skills, foreignlanguage aptitude, and foreign language grades among high-, average-, and low-proficiency foreign-language learners: two studies. *Language Testing* 15(2), 181–216.

Spyriadakis, J. H., & Standal, T. C. (1987). Signals in expository prose: Effects on reading. *Reading Research Quarterly, 22, 285-298.*

Staub, A., & Rayner, K. (2006). Eye movements and on-line comprehension processes. In M. G. Gaskell (Ed.), *Oxford Encylopedia of Psycholinguistics.* Oxford: Oxford University Press.

Stewart, M. R., & Grobe, C. H. (1979). Syntactic maturity, mechanics, and vocabulary and teachers' quality ratings. *Research in the Teaching of English, 13, 207-215.*

Stock, P. L., & Robinson, J. L. (1987). Taking on testing: Teachers as testers researchers. *English Education, 19,* 93-121.

Stuhlmann, J., Daniel C., Dellinger, A., Kenton, R., & Powers, T. (19999) A generlizability study of the effects of training on teachers' ability to rate children's writing using a rubric. Reading Psychology, Volume 20, Number 2, pp. 107-127(21)

Sturt, P. (2003). The time course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48: 542–562.

Sturt, P., & Lombardo,V. (2005). Processing coordinated structures: incrementality and connectedness. *Cognitive Science*, 29: 291–305.

Sullivan, F. J. (1987). Negotiating expectations: Writing and reading placement tests. Paper presented at the meeting of the Conference on College Composition and Communication, Atlanta.

Thorndyke PW, Hayes-Roth B 1979 The use of schemata in the acquisition and transfer of knowledge. *Cognitive Psychology, 11*:82106

Tierney, R. J., & Mosenthal, J. H. (1983). Cohesion and textual coherence. *Research in the Teaching of English, 17, 215-229.*

Trabasso, T., Secco, T., & van den Broek, P. (1984). Causal cohesion and story coherence. In H. Mandl, N. L. Stein, & T. Trabasso (Eds.), *Learning and Comprehension of Text* (pp. 83-111). Hillsdale, NJ: Lawrence Erlbaum.

Tryon, R.C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54*, 229-249.

Tung, P. (1986). Computerized adaptive testing: Implications for language test developers. In C.W. Stansfield (Ed.), *Technology and language testing (pp. 13-28*). Washington, DC: TESOL

van den Broek, P (1988).The effects of causal relations and hierarchical position on the importance of story statements. *Journal of Memory and Language, 27*, 1-22.

van den Broek, P., Tzeng, Y., Risden, K. Trabasso, T. & Bashe, P. (2001). Inferential questioning: Effects on comprehension of narrative texts as a function of grade and timing. *Journal of Educational Psychology*, *93*, (*3*) 521-529

van der Linden,W., & Glas, C. (Eds.). (2000). *Computer adaptive testing: Theory and practice.* Boston, MA: Kluwer Academic Publishers.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Vaughan, C. (1991). Holistic assessment: What does on in the rater's mind? In Hamp-Lyons, L., editor, *Assessing second language writing in academic contexts.* Norwood, NJ: Ablex, 111-25.

Veal, L. R. (1974). Syntactic measures and rated quality in the writing of young children. *Studies in Language Education, Report No. 8*. Athens: University of Georgia. (ERIC Document Reproduction Service No. 090 55).

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English, 17, 285-296*.

Weigle, S.C. (1994). Effects of training on raters of ESL compositions: quantitative and qualitative approaches. Unpublished PhD dissertation, University of California, Los Angeles.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing* 15, 263-87.

White, E. M. (1985). *Teaching and Assessing Writing*. San Francisco: Jossey-Bass.

Witte, S. P. (1983a). Topical structure and revision: An exploratory study. *College Composition and Communication, 34, 313-339.*

Witte, S. P. (1983b). Topical structure and writing quality: Some possible text-based explanations of readers' judgments of students' writing. *Visible Language, 17, 177-205.*

Witte, S. P., Daly, J. A., & Cherry, R. D. (1986). Syntactic complexity and writing quality. In D. A. McQuade (Ed.), *The Territory of Language (pp. 150-164).* Carbondale, IL: Southern Illinois University Press.

Witte, S. P., & Faigley, L. (1981). Coherence, cohesion and writing quality. *College Composition and Communication, 32, 189-204.*

Ziefle, M. (1998) Effects of display resolution on visual performance, *Human Factors, 40 (4), 555-568*

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 386-397.*

Zwaan, R. A., Radvansky, G. A., Hilliard, A. E., & Curiel, J. M. (1998). Constructing multidimensional situation models during reading. *Scientific Studies of Reading, 2, 199-220.*

APPENDIX A

EPT RATER SURVEY

Thank you for participating in the TOEFL iBT Writing Study. To help us improve our future efforts, please take a few minutes to complete this survey. We welcome any comments and suggestions you might offer.

Name: **_____**

1. Overall, were you satisfied with the qualities of the following aspects in rater training?

	Very Satisfied	Satisfied	Somewhat Satisfied	Not at all satisfied
Training Personnel	**O**	**O**	**O**	**O**
Facilities	**O**	**O**	**O**	**O**
Sample Rating Rubric	**O**	**O**	**O**	**O**
Rating Tour	**O**	**O**	**O**	**O**

2. When you grade a TOEFL essay, how important do you think the following factors are to successful essay writing? Please check the appropriate circle for each criterion.

	To a large degree	somewhat	To a small degree	Not at all
Organization	**O**	**O**	**O**	**O**
Development	**O**	**O**	**O**	**O**
Grammar and Lexical choice	**O**	**O**	**O**	**O**
Content (relevant to the given essay topic)	**O**	**O**	**O**	**O**
Plagiarism	**O**	**O**	**O**	**O**
Essay length	**O**	**O**	**O**	**O**
Sentence complexity	**O**	**O**	**O**	**O**

3. While you were rating a TOEFL essay, approximately how often did you refer to the scoring rubrics? Please check the appropriate circle.

	Never	Once or twice	3 to 5 times	More than 5 times
a. The scoring rubrics	**O**	**O**	**O**	**O**

4. After participating in the training session and rating TOEFL iBT essays, how confident did you feel about evaluating essays in each of the following criteria? Please check the appropriate circle.

	To a large degree	somewhat	To a small degree	Not at all
Organization	**O**	**O**	**O**	**O**
Development	**O**	**O**	**O**	**O**
Grammar and Lexical choice	**O**	**O**	**O**	**O**
Content (relevant to the given essay topic)	**O**	**O**	**O**	**O**
Plagiarism	**O**	**O**	**O**	**O**
Essay length	**O**	**O**	**O**	**O**
Sentence complexity	**O**	**O**	**O**	**O**

Please give us your opinions about the importance of various aspects of writing by checking the appropriate circle for the questions below.

5. In general, how important do you think the following factors are to successful essay writing? Check the appropriate circle for each dimension.

	To a large degree	somewhat	To a small degree	Not at all
Organization	**O**	**O**	**O**	**O**
Development	**O**	**O**	**O**	**O**
Grammar and Lexical choice	**O**	**O**	**O**	**O**
Content (relevant to the given essay topic)	**O**	**O**	**O**	**O**
Plagiarism	**O**	**O**	**O**	**O**
Essay length	**O**	**O**	**O**	**O**
Sentence complexity	**O**	**O**	**O**	**O**

6. In your own teaching, when you evaluate students' essays, how important are the following factors to the final grades you assign? Check the appropriate circle.

	To a large degree	somewhat	To a small degree	Not at all
Organization	**O**	**O**	**O**	**O**
Development	**O**	**O**	**O**	**O**
Grammar and Lexical choice	**O**	**O**	**O**	**O**
Content (relevant to the given essay topic)	**O**	**O**	**O**	**O**
Plagiarism	**O**	**O**	**O**	**O**
Essay length	**O**	**O**	**O**	**O**
Sentence complexity	**O**	**O**	**O**	**O**

Please tell us about your previous experiences evaluating writing by responding to the following questions.

7. In the past three years, have you engaged in any of the following assessment activities? Check the appropriate circle.

	Yes	No
a. Used a holistic rubric or scoring guide to evaluate writing?	**O**	**O**
b. Used an analytic or trait-based rubric/scoring guide to evaluate writing?	**O**	**O**

To help us describe the diverse backgrounds and experiences of raters who participated in this study, please answer the following questions.

8. Approximately how many years have you taught the following? Check the appropriate circle.

	None	1-3 years	4-6 years	7-9 years	10 or more
a. ESL/EFL (any type of class)	**O**	**O**	**O**	**O**	**O**
b. English composition/academic writing	**O**	**O**	**O**	**O**	**O**
c. Academic writing to ESL/EFL students	**O**	**O**	**O**	**O**	**O**
d. English Grammar	**O**	**O**	**O**	**O**	**O**

9. Comments? Suggestions? Ideas? Reflections? (Please write below.)

APPENDIX B

RATING RUBRICS FOR SEEPT COMPOSITION SCORING

Revised 07/07; Diana Xin Wang

Grade 1: Too low: Place in ESL 500 (identify for tutoring).

A. Organization

· Length insufficient to evaluate; (or)
· No organization of ideas

B. Development

· No cohesion, like a free writing;
· No support of elaboration of ideas
· Insufficient length to evaluate
· Irrelevant to assigned topic
· Completely lack of main idea

C. Grammar and Lexical Choice

· Grammar and lexical errors are severe;
· No sentence complexity
· Simple sentences are flawed

D. Plagiarism

· Majority of essay copied without documentation

Grade 2: ESL 500

A. Organization

· Length may be insufficient to evaluate;
· Elements of essay organization (intro, body and conclusion) may be attempted, but are simplistic and ineffective.

B. Development

· Essay may lack a central controlling idea (no thesis statement, or thesis statement flawed);
· Essay does not flow smoothly and ideas are difficult to follow
· Development of ideas is insufficient; examples may be inappropriate; logical sequencing may be flawed or incomplete
· Paragraph structure not mastered; lack of main idea (topic sentence), focus, and cohesion

C. Grammar and Lexical Choice

· Grammar and lexical errors impede understanding;
· Awkwardness of expressions and general inaccuracy of work forms
· Little sophistication in vocabulary and linguistic expression; little sentence variety; sentence complexity not mastered

D. Plagiarism

· Attempts at paraphrase are generally unskillful and inaccurate
· Some overt plagiarism

Grade 3: ESL 501

A. Organization

· Length is sufficient for full expression of ideas
· Elements of essay organization are clearly present, though they may be flawed

B. Development

· Attempt to advance a main idea; presence of thesis statement
· Flow somewhat smoothly
· Some development and elaboration of ideas; evidence of logical sequencing; transitions may show some inaccuracies
· Paragraph structure generally mastered, generally cohesive

C. Grammar and Lexical Choice

· Some grammatical/lexical errors; meaning may be occasionally obscured, but essay is still comprehensible
· Inconsistent evidence of some sophistication in sentence variety and complexity

D. Plagiarism

· Covert plagiarism; attempted summary and paraphrase; may contain isolated instances of direct copying; may not cite sources, or may cite them incorrectly
· Moderately successful paraphrase in terms of smoothness

Grade 4: Exempt from ESL 501

A. Organization

· Contain a clear intro, body and conclusion

B. Development

· Clear thesis statement, appropriately placed
· Good development of thesis; logical sequencing; reasonable use of transitions
· Paragraphs are fairly cohesive

C. Grammar and Lexical Choice

· May contain minor grammatical/lexical errors, but meaning is clear
· Strong linguistic expression exhibiting academic vocabulary, sentence variety and complexity

D. Plagiarism

· Effective, skillful summary and paraphrase
· Sources are cited, though possibly inaccurately

APPENDIX C

CONSENT FORM

Purpose and Procedures: This study is being conducted by Xin Wang and Dr. Fred Davidson in the Department of Educational Psychology, at the University of Illinois at Urbana-Champaign (UIUC). It is intended to look for the possible future revision of the ESL Placement Test scoring. If you agree to take part in this research, you will be asked to attend a 60-minute training session to learn how to use a computer-based rater interface and then grade 20 EPT writing samples on the interface. It takes approximately three hours for each rater to finish training and essay grading.

Voluntariness: Your participation in this research is voluntary. You may refuse to participate or withdraw your consent at any time and have the results of the participation removed from the experimental records. Your choice to participation or not will not affect your student status or your employment at this university.

Risks and Benefits: There is no more risk than what could be encountered in daily life. The experiment will not pose subject under any physical or psychological risk. Your participation may provide helpful information on the future application of computer-based rater interface in essay grading. A compensation of 50 US dollars will be paid to each participant after the experiment session.

Confidentiality: Only the researcher of this study will have access to research results associated with your identity. The dissemination of this investigation is the researcher's Ph. D dissertation, conference talks and possible publications. The results of this participation will be coded and dissemination will not contain any identifying information without the prior consent of the participant unless required by law.

Who to Contact with Questions: Questions about this research study should be directed to the researcher, Xin Wang (Diana) in the Department of Educational Psychology at UIUC. She can be reached at xinwang2@uiuc.edu, or 217-766-3680. Questions about your rights as a research participant should be directed to the UIUC Institutional Review Board Office at 333.2670; irb@uiuc.edu or the Bureau of Educational Research at 333-3023. You will receive a copy of this consent form.

I certify that I have read this form and volunteer to participate in this research study.

_____

(Print) Name

_____ Date: _____

Signature

APPENDIX D
GLOSSARY

Variables	Description
LPS	Letter-per-second reading rate
WPM	Word-per-minute rate
Vocab	The total number of vocabulary excluding stop words
Word	The total number of vocabulary including stop words
Sentence	The total number of sentence
Subsent	The total number of sub-sentences
Trancount	The total number of transitional words
Trantype	The total number of different transitional words
Freq	The weighted average word frequency in essays, with the weight defined as word frequency of the vocabulary from Brown Corpus
Category of Tran. Word	The types of transitional words and
Tran. Word	The total number of transitional words
Word Per. Sentence	The average number of words in a sentence for each essay